1965

A statistical outlier methodology for observed points and lines

Florence Gertrude Tetreault
Iowa State University

A STATISTICAL OUTLIER METHODOLOGY FOR OBSERVED POINTS AND LINES

by

Florence Gertrude Tetreault

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
The Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major Subject: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

Head of Major Department

Signature was redacted for privacy.

Dean of Graduate College

Iowa State University
Of Science and Technology
Ames, Iowa

1965

TABLE OF CONTENTS

		Page
I.	INTRODUCTION	1
II.	REVIEW OF LITERATURE	8
III.	ESTIMATION AND HYPOTHESIS TESTING SUBSEQUENT TO A PRELIMINARY TEST FOR A UNIVARIATE STATISTICAL OUTLIER	30
	A. Introduction	30
	B. Estimation and Hypothesis Testing Subsequent to an Outlier Test Assuming No A Priori Information	31
	1. Statement of the problem	31
	2. Thompson's criterion	35
	3. Estimation of μ_1 after Thompson's test for an outlying observation	35
	4. Rule of procedure	35
	5. Derivation of $E(\bar{x}^*)$	37
	6. Mean square error of $\bar{x}^*$	44
	7. Test for $H_0: \mu_1 = \mu_0$ after Thompson's test for an outlying observation has been made	46
	8. Power of the test procedure given in Part 7	47
	C. Estimation and Hypothesis Testing Subsequent to an Outlier Test Assuming A Priori Information Sufficient to Identify a Suspected Outlier	53
	1. Statement of the problem	53
	2. Estimation of μ_1 after a preliminary test for an outlying observation	55

Page

3. Derivation of $E(\bar{x}^*)$ and the mean square error of $\bar{x}^*$ (σ^2 known) 55

4. Derivation of $E(\bar{x}^*)$ and the mean square error of $\bar{x}^*$ (σ^2 unknown) 63

5. Test for $H_0: \mu_1 = \mu_0$ versus $H_a: \mu_1 \neq \mu_0$ (σ^2 known) 73

6. Power of the test procedure given in Part 5 74

7. Test for $H_0: \mu_1 = \mu_0$ versus $H_a: \mu_1 \neq \mu_0$ after a preliminary test for an outlying observation has been made (σ^2 unknown) 79

8. Power of the test procedure given in Part 7 80

IV. LINE OUTLIER THEORY WITH SLOPE AND INTERCEPT CONSIDERED SEPARATELY 87

A. Introduction 87

B. The Problem of Slope Outliers 89

C. Modifications of Point Outlier Criteria 90

1. Extreme deviation statistic 90

2. The statistic S_N^2/S^2 91

3. Dixon's statistic 92

D. The Problem of Intercept Outliers 92

E. Outlier Tests Assuming Available A Priori Information Sufficient to Identify a Suspected Slope Outlier 93

1. The construction of the statistic V 93

2. Numerical example 95

3. Derivation of the frequency function of V 96

Page

4. Calculation of the expected mean squares 97

5. Power of the test 99

6. Special case 101

7. The construction of the statistic V' 102

8. Numerical example 103

9. The distributional structure of V' 103

10. The approximate distribution of V' 104

11. Size of the test using the approximate distribution 105

V. LINE OUTLIER THEORY WITH SLOPE AND INTERCEPT CONSIDERED JOINTLY 107

A. Introduction 107

B. A Test Criterion Based on Maximum Distance 109

C. A Test Criterion Based on Wilk's Statistic 114

D. A Test Criterion Based on Siotani's Statistic 117

E. Empirical Comparison of the Criteria Discussed in Sections B, C and D 120

F. Outlier Tests Assuming Available A Priori Information Sufficient to Identify a Suspected Outlier 122

1. Construction of the U statistic 122

2. Derivation of the distribution of U 123

3. Distributional structure of U under the alternative hypothesis 125

4. Special case 126

5. Construction of the statistic U' 127

6. Distributional structure of the U' sta-

Page

tistic under the null hypothesis 129

7. Approximate distribution of U' 130

8. Size of the test using the approximate distribution 131

VI. SUMMARY 132

VII. LITERATURE CITED 134

VIII. ACKNOWLEDGEMENTS 138

I. INTRODUCTION

The problem of outlying observations has been the object of considerable investigation and, as in the case of many other problems resulting in theoretical developments, it arose from a practical need. Almost every scientist at some time is confronted with a set of data in which at least one observation differs radically from the remaining ones, and this may lead him to suspect that the series of observations does not represent a random sample from a single population. Consider the following illustrative examples:

Example 1. Chauvenet (1876) gave the following residuals of fifteen observations of the vertical diameter of Venus made by Lieutenant Hearndon in 1846: -0.30, -0.44, -0.24, -0.13, 0.06, -1.40, -0.22, -0.05, 0.20, 0.39, 0.48, 1.01, 0.63, 0.18, 0.10. The residuals 1.01 and -1.40 are much larger in absolute value than the others.

Example 2. K. Pearson (1931) gave the following capacities (in cubic centimeters) of seventeen male Moriori skulls: 1230, 1318, 1380, 1420, 1630, 1378, 1348, 1380, 1470, 1445, 1360, 1410, 1540, 1260, 1364, 1410, 1545. The highest value, 1630, appears to be anomalous.

Example 3. The following data was given by McKay (1935). In the course of routine testing of a standard leather product of a tannery five parallel tests yielded the values 32.44, 36.45, 39.64, 40.13 and 41.09, for the hide substance content of the

leather specimens. The first observation appears to be low and the question facing the tannery chemist is whether it is worth while instituting inquiries as to why such a low value occurred. Long experience with the product in question has established a value of 2.226 for the standard deviation.

Example 4. The following example is due to Quesenberry and David (1961). A sample of six observations was drawn from a table of random normal numbers and a randomly selected observation was increased by two standard deviations. The observations obtained were 265, 223, 291, 105, 43 and 477. Is the value 477 anomalous? A second sample of six observations was drawn from a table with the same variance but with a different mean to give an independent estimate of variance. These observations were 171, 111, 185, 68 and 217.

These four examples typify situations which may actually arise in scientific experimentation, and they emphasize the need for basic theoretical research on anomalous observations.

How could such discordant observations occur? First, the aberrant observation may, of course, be due to errors of measurement or recording, in which case, if such be known, it should be rejected. Second, the observations may all be from the same population in which case the observed difference is the chance result of taking a sample of limited size. Since all the observations in this case are valid, the scientist obtains biased results should he reject any observation.

Third, the aberrant observation may really indicate that the assumed model is incorrect, i.e. that the sample represents observations from populations that are not identical.

When the scientist knows that an abnormal error or blunder has been made, he would, of course, not hesitate to discard such an observation. When the scientist does not have enough practical grounds to support either accepting or rejecting an extreme observation, he must resort to some kind of statistical judgement. He would like to answer the question; "What is the probability that the observed differences are due solely to random sampling errors?", in such a way that there is little doubt that certain observations should be rejected.

The approach to the problem of outlying observations depends upon the object in mind. If one is solely interested in determining whether an observation is an outlier, in order, perhaps, to investigate the condition or conditions that may have led to this extreme observation, then the test for such an outlying observation is an end in itself. If, on the other hand, one is interested in pruning the observations in order to obtain a more accurate estimate of some population parameter, say the population mean, then one is interested not only in a test for an outlying observation, but also in the estimation of the parameter subsequent to the outlier test. Thus one would also consider the possible bias of the estimate and its mean square error, taking proper account of the use of the

outlier test. If the sample data, subsequent to an outlier test, is to be used to test hypotheses about a population parameter, then one is interested not only in a criterion for an outlier but also in the power and size of subsequent tests of hypotheses. In each of the last two cases, the test for an outlying observation is not an end in itself, and could therefore be termed a preliminary test. It is our contention that since the usual purpose of obtaining a sample is to estimate a population parameter or to test hypotheses about a population parameter, we should be concerned with how well this is accomplished when the sample at hand contains a suspected outlier. Therefore, in Chapter III, we consider the outlier test to be a preliminary test and formulate the problem of point outliers in such a manner that the theory of incompletely specified models may be applied.

The classical method of handling the problem of detecting a point outlier is to assume that the sample observations come from a normal population, devise an appropriate outlier test statistic, derive the distribution of this test statistic under the null hypothesis that all the observations come from the same normal population, and then reject the hypothesis if the calculated test statistic for it is unlikely to have occurred in random sampling. The usual test statistic, formulated on the assumption that the scientist looks at the sample results of an experiment and then notes that he has a discord-

ant observation, involves the ratio of the difference between the extreme value and the mean value, and either the population standard deviation or an estimate of it obtained either from the sample at hand or from an independent sample or a combination of these two. This statistic is referred to as the extreme deviate statistic. A survey of the literature on outliers is given in Chapter II. All of the studies mentioned therein are concerned only with the problem of identification of outliers. Anscombe (1960) discussed the problem of subsequent estimation in a general way but did not give any specific results as to the possible bias and size of the mean square error of the estimate obtained subsequent to an outlier test.

In Chapter III, as we mentioned previously, we are concerned with the estimation problem and the problem of testing hypotheses when the scientist does not know in advance that his sample may contain observations from two different populations. In other words, the test for an outlying observation is considered to be a preliminary test and we deal with the two subsequent problems:

(i) the estimation of the population mean on the basis of the outlier test, and

(ii) the size and power of subsequent tests of hypotheses concerning the population mean.

In contrast to the above situation, suppose that the

scientist knows in advance that his sample may contain observations from two populations, say two normal populations with the same variance but with different means. For example, geologists know that sometimes large boulders are observed in gravel deposits, and their occurrence raises the question: "Are such large boulders part of the pebble population, or are they outliers transported by extraneous agents such as ice?" Or suppose that the scientist does not have complete control over his measuring devices, i.e. he knows that his measuring process is subject to erratic behaviour, and before proceeding with the experiment suspects that this might result in sample data from two normal populations with equal variances but unequal means. We could also envisage a situation where the scientist has some a priori information of a kind that makes it possible for him to say that any observation in a sample would be suspect if it were greater than C (or less than C), where C is known from a priori information. This problem is also considered in Chapter III.

In Chapters IV and V of this thesis the univariate situation is extended to the bivariate situation, i.e. we consider the construction of an outlier methodology for straight lines. It is sometimes necessary to know whether several regression lines obtained from scientific experiments are parallel. For example, the usual kind of covariance analysis assumes that the slopes of the regression lines are the same, i.e. that the

lines are parallel. The purpose of the covariance analysis is to determine whether these parallel lines are significantly displaced from one another as a result of the treatments. As a second example, consider biological assays which are based on parallel lines. The regression line of response on dosage is determined for each compound of unknown strength and for a known standard. The distance between two parallel lines gives the relative potency of the two compounds. However, interpretation is impossible when the lines are not parallel.

In some situations it is also of importance to know whether the population regression equations are identical even to the constant term. If it be true that they are identical, the scientist may want to obtain an estimate of the single population regression line by combining the results of several experiments. In Chapter IV several criteria for testing the hypothesis that a particular slope (either the largest or the smallest) is not an outlier are proposed, and in Chapter V several somewhat different approaches are made to find a statistic to test the hypothesis that an entire line is not an outlier line.

The final chapter summarizes the results of this thesis.

II. REVIEW OF LITERATURE

We now take a small side trip through the past to see how the problem of outlying observations was handled by various authors. In general their papers can be divided into two classes; those dealing with a single sample from a one-dimensional normal population with unknown parameters and those dealing with samples from one-dimensional normal populations for which one or both of the parameters are known or are estimated from independent samples. In practice, the mean and variance of a population are not generally known and must be estimated from the sample itself.

In this chapter, the criteria proposed by the various authors are illustrated by the examples given in Chapter I. For the sake of completeness, the criteria are applied as they would have been when they were first proposed. For example, some of the criteria proposed assume a known population variance but were applied to problems where the population variance was unknown and was estimated from the sample at hand. Note that the variance of the population is unknown in Examples 1 and 2, is known in Example 3, and is estimated from an independent sample in Example 4.

Pierce (1852) is credited with being the first to propose a criterion for the rejection of an outlier. He said:

> Observations should be rejected when the probability of the system of errors obtained by retaining them

> is less than that of the system of errors obtained by their rejection multiplied by the probability of making so many and no more abnormal observations. In determining the probability of these two systems of errors, it must be carefully observed that, because observations are rejected in the second system, the corresponding observations in the first system must be regarded, not as being limited to their actual values, but only as surpassing the limit of rejection.

He used the following notation:

n = total number of observations,

N = number of observations to be rejected,

n' = number of observations to be retained,

$x_1, x_2, \ldots, x_n$ = system of errors when all observations are retained,

$x'_1, x'_2, \ldots, x'_{n'}$ = system of errors when N observations are rejected,

σ = standard deviation of the first system,

σ' = standard deviation of the second system.

Then

$$2\int_{k\sigma}^{\infty} \phi(x)\,dx = (2/\sqrt{2\pi}\,\sigma)\int_{k\sigma}^{\infty} e^{-x^2/2\sigma^2}\,dx = \psi(k)$$

is the probability of an error numerically greater than $k\sigma$, and

$$P = \phi(x_1)\,\phi(x_2)\ldots\phi(x_n)\,(dx)^n\,[\psi(k)]^N/[\phi(k\sigma)\,dx]^N$$

$$= (2\pi\sigma^2)^{-n'/2}\; e^{-(n-1-Nk^2)/2}\; (dx)^{n'}\; [\psi(k)]^N$$

is the probability of the first system of errors, using the condition that N of them exceed k . The probability of the second system multiplied by the probability that only the N

outliers were subjected to a disturbing influence is given by

$$P_1 = \phi_1(x_1') \, \phi_1(x_2') \ldots \phi_1(x_{n'}') \, (dx)^{n'} \, y^N \, (1-y)^{n'}$$

$$= (\sqrt{2\pi}\,\sigma')^{-n'} \, e^{-(n'-1)/2} \, (dx)^{n'} \, y^N \, (1-y)^{n'},$$

where y is the probability that a disturbing influence caused such an unusual observation that it is rejected. In order to reject the N observations, Pierce said that P must be less than P_1, i.e.

$$(\sigma'/\sigma)^{n'} \, e^{N(k^2-1)/2} \, [\psi(k)]^N < y^N \, (1-y)^{n'}.$$

The value of y must be determined by using the condition that P_1 be a maximum and the value of k must be found by a series of approximations. Stewart (1920) pointed out that Pierce's criterion was incorrect for $N > 1$ and Airy (1856) gave an example where the use of Pierce's criterion led to poor results.

Another criterion was given by Stone (1867). He introduced his criterion in these words:

> I assume that a particular person, with definite instrumental means and under given circumstances is likely to make, on the average, one mistake in the making and registering of m observations of a given class.

He called m the modulus of carelessness and wrote

$$(2/\sqrt{2\pi}\,\sigma) \int_{k\sigma}^{\infty} e^{-x^2/2\sigma^2} \, dx = \psi(k) = 1/m.$$

Stone then said that all deviations that are greater in absolute value than $k\sigma$ are with greater probability to be attrib-

uted to mistakes rather than to ordinary errors, and therefore, the corresponding observations should be rejected.

Czuber (1891) objected to Stone's criterion and used the following example to support his argument. If we assume that a mistake is made once in 100 observations ($m = 100$), then all observations that deviate from the mean by more than 2.58σ will be discarded. However, if two hundred observations are made, then by the normal probability law one observation would lie beyond $\pm 2.81\sigma$ and, by Stone's criterion, would be rejected.

Chauvenet (1876) gave a criterion for the rejection of one observation. He said:

> We have seen that the function $(2h/\sqrt{\pi})\int_0^{a\rho/r} e^{-h^2\Delta^2} d\Delta$ represents in general the number of errors less than a which may be expected to occur in any extended series of observations when the whole number of observations is taken as unity, r being the probable error of an observation. If this be multiplied by the number of observations, n, we shall have the actual number of errors less than a; and hence the quantity
>
> $$n - n\Phi(t) = n[1 - \phi(t)]]$$
>
> expresses the number of errors to be expected greater than the limit a. But if this quantity is less than 1/2 it will follow that an error of magnitude a will have a greater probability against it than for it, and may, therefore, be rejected.

Therefore, he rejects those errors for which

$$n[1 - \phi(t)] < 1/2$$

or

$$\phi(t) > (2n-1)/2n.$$

In modern notation, Chauvenet's criterion would then reject observations that are numerically greater than $k\sigma$, where

$$(2/\sqrt{2\pi}\ \sigma) \int_0^{k\sigma} e^{-x^2/2\ \sigma^2}\ dx = (2n-1)/2n.$$

Chauvenet's argument has been severely criticized. Many writers deem it incorrect because $n[1 - \phi(t)]$ is a frequency and not a probability. Objections have also been raised concerning Chauvenet's choice of the value 1/2. In effect, he said that on the average a mistake occurs once in 2n observations, where n is the number of observations in the sample under consideration. It should be noted that Chauvenet's criterion is the same as Stone's with m=2n.

To illustrate the use of Chauvenet's criterion we use Examples 1, 2 and 3 given on pages 1 and 2. Example 4, in which an estimate of the population variance is available from a sample independent of the one containing the aberrant observation, will be used later to illustrate criteria proposed to handle this situation.

Example 1. The residuals, 1.01 and -1.40, are much larger numerically than the others. In order to apply the criterion, we calculate s = 0.5326 and (2n-1)/2n = 0.9667. Chauvenet's criterion would then reject observations that are numerically larger than $k\sigma$, where

$$(2/\sqrt{2\pi}\ \sigma) \int_0^{k\sigma} e^{-x^2/2\sigma^2}\ dx = 0.9667, \qquad (1)$$

and $s = 0.5326$ is used as an approximation for σ. Let $y = x/\sigma$, then (1) becomes

$$(1/\sqrt{2\pi})\int_0^k e^{-y^2/2}\,dy = 0.4834, \tag{2}$$

and we have $k = 2.13$ and $k\sigma = 1.13$. Chauvenet's criterion would then reject observations that are numerically larger than 1.13, and hence the observation corresponding to the deviation -1.40 is rejected. Consider the remaining fourteen observations. We have $s = 0.4048$ and $(2n-1)/2n = 0.9643$. From

$$(2/\sqrt{2\pi}\,\sigma)\int_0^{k\sigma} e^{-x^2/2\sigma^2}\,dx = 0.9643, \tag{3}$$

we find $k\sigma = 0.85$ and hence we would reject the value 1.01. Applying the criterion to the remaining observations, we find that no more can be rejected.

Example 2. The mean and standard deviation of the sample observations are 1405.2 and 97.83, respectively. From

$$(2/\sqrt{2\pi}\,\sigma)\int_0^{k\sigma} e^{-x^2/2\sigma^2}\,dx = 0.9706,$$

we obtain $k = 2.178$ and $k\sigma = 213.07$. According to this, we reject observations that are greater than 1405.2 + 213.07 or less than 1405.2 - 213.07. Thus Chauvenet's criterion leads us to the rejection of 1630.

Example 3. The mean of the sample is 37.95 and the problem stated that long experience with the product has established

a value of 2.226 for the standard deviation. Then from

$$(2/\sqrt{2\pi}\,\sigma)\int_0^{k\sigma} e^{-x^2/2\sigma^2}\,dx = .9,$$

we obtain $k = 1.645$ and $k\sigma = 3.66$. Observations larger than $37.95 + 3.66$ or smaller than $37.95 - 3.66$ would be rejected. Therefore, the observation 32.44 is rejected.

Irwin (1925) proposed a statistic based upon the fact that if the observations be arranged in order of magnitude, it is possible to obtain the frequency distribution of the difference between the pth and the (p+1)th observations. The criterion λ is $1/\sigma$ times the interval between successive observations arranged in descending order of magnitude. Thus the test for a single outlier x_n (or x_1) is based on the statistic $\lambda = (x_n - x_{n-1})/\sigma$ [or $(x_1 - x_2)/\sigma$]. If there be k large outliers, the test would be based on $(x_{n-k+1} - x_{n-k})/\sigma$ and for k small outliers, $(x_{k+1} - x_k)/\sigma$. The application of the criterion is simplified by two tables given by Irwin. Table II gives values for $P_1(\lambda)$, the probability that the first and second observations from either end should differ by more than λ times the standard deviation of the population and Table III gives values for $P_2(\lambda)$, the same function for the second and third observations from either end. In actual practice, σ is replaced by its estimated value from the sample. However, as Irwin points out, when the sample size is small the standard deviation of the sample is a very unreliable measure of the

standard deviation of the population. For illustration, consider the examples given previously.

Example 1. Arranging the fifteen observations in descending order of magnitude we have 1.01, 0.63, 0.48, 0.39, 0.20, 0.18, 0.10, 0.06, -0.05, -0.13, -0.22, -0.24, -0.30, -0.44, -1.40. The difference between the first and second observation is 0.38 and we have $\lambda = 0.38/0.5326 = 0.713$. Using Table II in Irwin's paper, we find $P_1 = 0.241$. The difference between the last two observations is 0.96 and $\lambda = 0.96/0.5326 = 1.802$. From Table II we have $P_1 = 0.014$. This indicates that -1.40 should be rejected and 1.01 should not.

Example 2. In this problem $\lambda = (1630 - 1545)/97.83 = 0.869$. Referring to Table II, we find $P_1 = 0.166$, and Irwin's criterion would not reject the observation 1630.

Example 3. We have $\lambda = (36.45 - 32.44)/2.226 = 1.801$ and $P_1 = 0.076$. Irwin's criterion would not reject the observation 32.44.

Another method suggested by Irwin (1925) is to find from the ordinary tables of the probability integral the probability of an observation so divergent as the outlying one occurring at all. He said:

> For example suppose we find in a series of 1000 observations one which is greater than the mean by 3.5 times the standard deviation and that the one before it is greater than the mean by 3.0 times the standard deviation. We find from the tables of the probability integral the chance of an observation occurring

> more distant from the mean than the mid-point between these two: that is, 3.25 times the standard deviation from the mean. That is, 0.0006, or in 1000 observations we should expect 0.6 of an observation to be so distant; and we should not be justified in rejecting it. But if the outlying observation were four times the standard deviation from the mean, noting that the probability of a deviation greater than 3.5σ is 0.0002 and greater than 3.75σ is 0.0001, we should certainly reject an observation whose deviation from the mean was 4.5σ seeing that the probability of a deviation greater than 4.0σ is 0.00003.

Let us apply this idea to our examples.

Example 1. Although the observations have been referred to as "residuals", their sum is 0.27, so that the sample mean is 0.018. First consider the observation 1.01. We must calculate (1.01 + 0.63)/2 = 0.82. Then Z_1 = (0.82 - 0.018)/0.5326 = 1.506. Since $P(Z > Z_1) = 0.066$ and 15(0.066) = 0.990, we are not justified in rejecting the observation. Now consider the observation -1.40. We have (-1.40 - 0.44)/2 = -0.920 and Z_1 = (-0.920 - 0.018)/0.533 = -1.761. Since $P(Z < -1.761)$ is equal to 0.039 and 15(0.039) = 0.585, we are not justified in rejecting the observation -1.40.

Example 2. In this case (1630 + 1545)/2 = 1587.5 and Z_1 = (1587.5 - 1405.2)/97.83 = 1.863. Then $P(Z > 1.863)$ is equal to 0.031 and 17(0.031) = 0.527. Therefore, we cannot reject the observation.

Example 3. We have (32.44 + 36.45)/2 = 34.445 and hence the value of Z is given by Z_1 = (34.445 - 37.95)/2.226 = -1.575. Since $P(Z < -1.575)$ is equal to 0.058 and 5(0.058) = 0.290, we

cannot reject the observation.

The preceding criterion is very similar to one proposed by Wright (1884). He suggested rejecting any observation whose residual exceeds five times the standard deviation, because if the normal law is satisfied, only about one observation in 1000 would be rejected.

Tippett (1925) investigated the possibility of using the range to determine whether an outlying member of a sample should be rejected. In symbols his criterion is $W/\sigma = (x_n - x_1)/\sigma$. He gave tables of the mean range expressed in terms of the population standard deviation, σ, for sample sizes n=2 to 1000. Later Pearson (1932), using the earlier work of Tippett, constructed a table of percentage limits for the distribution of the range in samples from a normal population. We now apply Tippett's criterion to Examples 1, 2 and 3.

Example 1. The range is 1.01 + 1.40 = 2.41. From Pearson's Table A, we find that the 5% limit of the range for a sample of size fifteen is 4.79(0.5326) = 2.551, where s = 0.5326 is used as an estimate of σ. According to Tippett's criterion we would not reject any observation because the observed range is less than 2.551.

Example 2. The highest and lowest observations are 1630 and 1230, giving a range of 400. For a sample of size seventeen from a distribution with σ estimated by 97.83, Table A shows

that the 5% limit of the range is 4.89(97.83) = 478.479. The observed range is less than this, so all the observations are retained.

Example 3. In this case the range is 41.09 - 32.44 = 8.65, and the 5% limit of the range for samples of size five is 3.87(2.226) = 8.615. We conclude that the observation 32.44 is an outlier.

McKay (1935) suggested using the statistic $u = (x_n-\bar{x})/\sigma$ if the largest observation were suspect and $u = (\bar{x}-x_1)/\sigma$ if the smallest observation were suspect. He derived the probability distribution for the extreme deviate, $x_n-\bar{x}$ (or $\bar{x}-x_1$), assuming a population variance of unity. It can be written explicitly only for n=2. In this case we have

$$f_2(u) = \frac{2}{\sqrt{\pi}} e^{-u^2}, \quad u > 0$$

which is the right half of the normal distribution. For n=3 McKay obtained

$$f_3(u) = \frac{3\sqrt{3}}{\pi} e^{-3u^2/4} \int_0^{3u/2} e^{-t^2} dt,$$

and for $n > 3$, he suggested using the approximation

$$P_n(u) = \frac{n}{\sqrt{2\pi}} \int_{u\sqrt{n/(n-1)}}^{\infty} e^{-t^2} dt.$$

Nair (1948) tabulated the exact values of $P_n(u)$ and showed that the approximation proposed by McKay was in good agreement with the exact values, and Grubbs (1950) simplified

the derivation of $P_n(u)$. The following illustrate McKay's criterion:

Example 1. We have n = 15, s = 0.5326 and $\bar{x}$ = 0.018. To test the observation 1.01, we calculate u = (1.01 - 0.018)/0.5326. We obtain u = 1.8626. Then $u\sqrt{15/14}$ = 1.928 and we find

$$P_{15}(u) = \frac{15}{\sqrt{2\pi}} \int_{1.928}^{\infty} e^{-t^2/2}\, dt = 0.4035.$$

We conclude that 1.01 is not an outlier. Now test the observation -1.40. We have u = (0.018 + 1.40)/0.5326 = 2.6624 and $u\sqrt{15/14}$ = 2.755. We obtain $P_{15}(u)$ = 0.045 and conclude that -1.40 is an outlier.

Example 2. Applying McKay's criterion to the data, we have u = (1630 - 1405.2)/97.83 = 2.2979. Then $u\sqrt{17/16}$ = 2.369 and $P_{17}(u)$ = 0.1513. We conclude that 1630 is not an anomalous value.

Example 3. In this problem u = (37.95 - 32.44)/2.226 = 2.4753, $u\sqrt{5/4}$ = 2.767 and $P_5(u)$ = 0.0140. We conclude that the observation 32.44 is an outlier, since the probability of getting a value 2.4753 or larger for u is 0.014.

Thompson (1935) suggested a criterion depending only upon the assumption of random sampling from a normal population. He showed that for a sample of size n with mean $\bar{x} = \sum_{i=0}^{n} x_i/n$, and variance $s^2 = \sum_{i=1}^{n} (x_i-\bar{x})^2/n$, the distribution of

$$\tau_i = (x_i-\bar{x})/s,$$

where x_i is any arbitrarily selected observation from the sam-

ple, is given by

$$p(\tau) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{(n-1)\pi}\ \Gamma\left(\frac{n-2}{2}\right)}\left(1 - \frac{\tau^2}{n-1}\right)^{(n-4)/2},$$

and that the frequency function of $t = (\tau\sqrt{n-2})/(n-1-\tau^2)$ is given by Student's t distribution with n-2 degrees of freedom.

Thompson proposed that for a given α, values of x_i for which $|\tau| > \tau_\alpha$ be rejected, i.e. considered outliers, where τ_α is chosen so that for any i, $P(|\tau| > \tau_\alpha) = \alpha$. He determined τ_α for $\alpha = \phi/n$, where $\phi = 0.05, 0.10, 0.20$ and $n = 3(1)22, 32, 42, 102, 202, 1002$. His criterion differs from those previously proposed in that it does not require that σ^2 be known. However, it should be noted that Thompson's criterion refers to an arbitrary observation and not to the smallest or largest observation in a sample. As Grubbs (1950) pointed out, the problem of finding the probability that an arbitrary observation is an outlier differs from the problem of finding the probability that a particular observation (say the largest) will be an outlier with respect to the remaining observations.

Pearson and Sekar (1936), investigating the same criterion in more detail, pointed out that it is only effective when there is a single outlier. A further discussion of their investigation of Thompson's criterion is given in Chapter III. For the present we simply apply Thompson's criterion to the examples.

Example 1. To apply Thompson's criterion to the observation -1.40, we find $\tau = (-1.40 - 0.018)/0.5326 = -2.662$ and $t = -2.662\sqrt{15-2}/\sqrt{15-1-(2.662)} = -3.65$. The tabular value of t for 13 degrees of freedom at the 5% level is 2.160, and Thompson's criterion rejects the observation. Consider the remaining fourteen observations. We have $s = 0.4048$ and $\bar{x} = 0.119$. To test the observation 1.01, we calculate $\tau = (1.01 - 0.119)/0.4048 = 2.201$ and $t = 2.67$. The tabular value of t at the 5% level with 12 degrees of freedom is 2.179, so Thompson's criterion also rejects this observation.

Example 2. In this problem $\tau = (1630 - 1405.2)/97.83$ and $t = 2.71$. Since $t_{0.05,15} = 2.131$, we see that Thompson's criterion rejects this observation.

Nair (1948) investigated the statistic $u = (x_n - \bar{x})/\sigma$ [or $u' = (\bar{x} - x_1)/\sigma$] first proposed by McKay (1935), obtained the distribution of u by a more direct method than that employed by McKay, and tabulated the probability integral for sample sizes n = 3(1)9. Whenever the population standard deviation is unknown, Nair suggested that it be estimated from another independent sample. He called this estimate s_ν (with ν degrees of freedom) and considered the distribution of the studentized form $t_n = (x_n - \bar{x})/s_\nu$ [or $(\bar{x} - x_1)/s_\nu$]. Percentage points of the distribution of t_n for sample sizes n = 3(1)9 and degrees of freedom ν = 10(1)20, 24, 30, 40, 60, 120, ∞ were obtained by Nair, and percentage points of t_n for n = 3,

4, 5 and $\nu \leq 10$ were obtained by Pillai (1959).

We now apply Nair's criterion to Example 4, which provides an independent estimate of the variance.

Example 4. The mean of the first set of observations is 234, and an independent estimate of the standard deviation, s_ν = 72.827, is obtained from the second set of observations. Then $t_n = (477 - 234)/72.827 = 3.337$. Using the table given by Pillai, we find that the 5% critical point for $n = 6$, $\nu = 5$ is 3.15 and conclude that the observation 477 is an outlier.

Tukey (1949) suggested a simple approximation to Nair's statistic t_n. He showed that

$$V = (u - 1.2\log_{10}n)/(0.75 + 3\nu^{-1}), \quad n > 3$$

or

$$V = (u - 0.5)/(0.75 + 3\nu^{-1}), \quad n = 3$$

may be treated as unit normal deviates. Tukey's approximation may be applied to Example 4 as follows:

Example 4. $V = (3.337 - 1.2 \log 6)/(0.75 + 0.6) = 1.78$. The observation 477 is not considered an outlier because the tabular value for the unit normal corresponding to the 5% level of significance is 1.96.

Kudô (1956) proposed a criterion to handle the situation where information about the population mean and variance is available from independent samples. He showed that the Pearson-Sekar statistic is a special case of his statistic and that it is optimum whereas Nair's statistic is not.

A criterion involving the ratio of the sums of squares of deviations from the mean for the truncated and for the complete sample was given by Grubbs (1950). The statistic proposed to test the significance of the largest (or smallest) observation is S_n^2/S^2 (or S_1^2/S^2), where S_n^2 is the sum of squares of the n-1 observations with the suspected outlier omitted, and S_1^2 is the sum of squares of the n observations. The statistic proposed to test the significance of the two largest (or two smallest) observations is $S_{n-1,n}^2/S^2$ (or $S_{1,2}^2/S^2$), where $S_{n-1,n}^2$ is the sum of squares of the n-2 observations with the two aberrant values omitted. The hypothesis is rejected whenever the value of the test statistic is too small to be accounted for by random sampling variation. Grubbs derived the probability distribution for his statistics and gave tables of percentage points; however, he did not discuss the power of the tests nor did he discuss the problem of one extreme at either end. Dixon (1953) recommended $(x_n-x_1)/s_\nu$ as a suitable criterion to test for one extreme at either end, and Pearson and Hartley (1943) prepared tables giving percentage points for $(x_n-x_1)/s_\nu$, where s_ν is estimated from a sample independent of the one under consideration.

Grubbs showed that $S_n^2/S^2 = 1 - (x_n-\bar{x})^2/(n-1)s^2$, or $S_n^2/S^2 = 1 - T_n^2/(n-1)$, where $s^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2/n$ and T_n is the studentized extreme deviate. He derived the exact distribution of the studentized extreme deviate and calculated probability

points. His alternative statistic S_n^2/S^2 has one advantage over the simpler statistic T_n; it permits the generalization mentioned previously. That is, a test for the second extreme is obtained by using the ratio of the sample variance of the first or last n-2 observations and the total sample variance.

Li (1964) showed that if the n observations are divided into two samples with the outlying observation in one and the remaining n-1 observations in another, then the reciprocal of Grubbs' criterion may be written as

$$S^2/S_n^2 = \frac{\text{total SS}}{\text{within-sample SS}}$$

$$= \frac{\text{among-sample SS + within-sample SS}}{\text{within-sample SS}}$$

$$= 1 + \frac{\text{among-sample SS}}{\text{within-sample SS}}$$

or

$$(n-2)\left[\frac{1}{S_n^2/S^2} - 1\right] = \frac{(\text{among-sample SS})/1}{(\text{within-sample SS})/(n-2)}$$

$$= \tilde{F}(1,n-2),$$

where $\tilde{F}(1,n-2)$ is the variance ratio with 1 and n-2 degrees of freedom. However, Li points out that $\tilde{F}$ does not follow the F distribution because the difference between the two sample means is made as large as possible.

To illustrate Grubbs' criterion we again use Examples 1 and 2.

Example 1. Using all fifteen observations we find that S^2 =

4.2496. Omitting the suspected outlier, -1.40, and using the remaining fourteen observations we have $S_1^2 = 2.0953$. Then $S_1^2/S^2 = 0.4931$. From the table in Grubbs' paper we find that $0.01 < P < 0.025$, so we would reject the observation at the 2.5% level of significance. We are now left with a sample of size fourteen. To test the significance of the observation 1.01, we calculate $S^2 = 2.0953$ and $S_n^2 = 1.2409$. Then $S_n^2/S^2 = 0.5922$ and we find that $0.05 < P < 0.10$. Therefore, Grubbs' criterion would not reject the observation 1.01.

Example 2. In this problem $S_n^2/S^2 = 0.5852$, where S^2 was calculated using the entire sample of seventeen observations and S_n^2 was calculated by omitting the anomalous value, 1630. We find that $0.025 < P < 0.05$ and reject the observation at the 5% level.

Dixon (1950) suggested that a ratio of ranges and subranges be used as a criterion for the rejection of outlying observations. The observations are arranged in ascending order if the suspected outlier is the smallest observation and in descending order if the largest observation is suspect. The criteria suggested depend on the size of the sample and are as follows:

$$r_{10} = (x_2-x_1)/(x_n-x_1), \quad \text{if } n = 3 \text{ to } 7,$$

$$r_{11} = (x_2-x_1)/(x_{n-1}-x_1), \quad \text{if } n = 8 \text{ to } 10,$$

$$r_{20} = (x_3-x_1)/(x_{n-1}-x_1), \quad \text{if } n = 11 \text{ to } 13,$$

$$r_{22} = (x_3 - x_1)/(x_{n-2} - x_1) \quad \text{if } n = 14 \text{ to } 30.$$

Critical values for the test statistics are given by Dixon and Massey (1951) in Table 8. Applying Dixon's criterion to Examples 1 and 2, we have:

Example 1. First consider the observation 1.01. The fifteen observations arranged in descending order of magnitude are 1.01, 0.63, 0.48, 0.39, 0.20, 0.18, 0.10, 0.06, -0.05, -0.13, -0.22, -0.24, -0.30, -0.44, -1.40. Since n = 15, we use r_{22} defined above. We have r_{22} = (0.48 - 1.01)/(-0.30 - 1.01) = 0.405. From Table 8 in Dixon and Massey (1951), we find that the probability of getting a value of 0.405 or larger for r is greater than 0.05. Therefore, we would not reject the observation 1.01. Now consider the value -1.40. Rearranging the observations, we have x_1 = -1.40 and x_{15} = 1.01. Then r_{22} = 0.585. Using Table 8, we find the critical value for fifteen observations at the 5% level is 0.525 and reject the observation.

Example 2. Ordering the observations in descending order and calculating r_{22}, we find that r_{22} = 0.289. We see from Table 8 that the observation 1630 is not rejected because the critical point for n = 17 at the 5% level of significance is 0.490.

In addition to proposing the ratio criterion mentioned above, Dixon also studied empirically the power functions of all the criteria discussed so far. As alternative hypotheses he assumed that the outliers were from normal distributions of

the form $N(\mu+\lambda\sigma,\sigma^2)$ or $N(\mu,\lambda^2\sigma^2)$. The power comparisons were based on sampling experiments with between 66 and 200 sets of observations in each case.

Assuming the existence of an independent estimate, s_ν^2, of σ^2, Quesenberry and David (1961) proposed a statistic to detect an outlying observation at one specified end of a sample and a statistic for two-sided testing. The assumption of an independent estimate of σ^2 was also made by Nair (1948). However, unlike Nair, Quesenberry and David also made use of the variance estimate, s^2, from the sample under consideration. To test for one outlier they suggested the statistic $b = (x_{max.}-\bar{x})/S$, where $S^2 = (n-1)s^2 + \nu s_\nu^2$. For two-sided testing, the statistic $b^* = \max_i |(x_i-\bar{x})/S|$ was proposed. Thus, if we denote by $y_1,y_2,\ldots,y_n$ the ordered values of the observations $x_1,x_2,\ldots,x_n$, $b^* = \max [(y_n-\bar{y})/S, (\bar{y}-y_1)/S]$, where the statistics in the brackets are the two possible extreme deviates in the sample. The distributions of both b and b^* were obtained and tables given for testing. To illustrate the method, we apply the criterion to Example 4.

Example 4. The sum of squares about the mean for the first set of observations is 116,504 and for the second set of observations is 26,519. Therefore, $S^2 = 143,023$ and we have $b = (477 - 234)/\sqrt{143,023} = 0.643$. The table in Quesenberry and David's paper gives the 5% point for $n = 6$ and $\nu = 5$ as approximately 0.638, so the observation 477 is rejected at the

5% level.

All of the papers mentioned so far deal with the problem of testing extreme observations arising in a sample from a one-dimensional normal distribution. Wilks (1963) considered the problem of identifying and testing extreme observations in a sample of size n from a multivariate normal distribution with unknown parameters. He considered the problem in detail for sets of 1, 2, 3 and 4 outliers and, although he did not find exact distributions of his criteria, he did present tables giving upper bounds for the amount of probability in the lower tail of the distribution of the test criterion for one outlier and the test criterion for a pair of outliers. A comparison of these upper bounds for a sample from a one-dimensional normal distribution with the exact values available from Grubbs (1950) tables for certain value of α is given by Wilks. No tables are given for the problem of identifying and testing three or more extreme observations. No attempt was made to study the power of the test criteria proposed. The problems of estimation and hypothesis testing subsequent to a multivariate outlier test were also not considered.

To summarize the discussion of the present chapter, we present in Table 1 the results obtained by applying the various criteria to Examples 1, 2, 3 and 4.

Table 1. Summary of numerical examples

	Example 1		Example 2	Example 3	Example 4
Observation	1.01,	-1.40	1630	32.44	477
Author					
Chauvenet	O[a]	O	O	O	D[b]
Irwin (first method)	N[c]	O	N	N	D
Irwin (second method)	N	N	N	N	D
Tippett	N	N	N	O	D
McKay	N	O	N	O	D
Thompson	O	O	O	D	D
Nair	D	D	D	D	O
Tukey	D	D	D	D	N
Grubbs	N	O	O	D	D
Dixon	N	O	N	D	D
Quesenberry and David	D	D	D	D	O

[a]The observation is an outlier.

[b]The criterion does not apply.

[c]The observation is not an outlier.

III. ESTIMATION AND HYPOTHESIS TESTING SUBSEQUENT TO A PRELIMINARY TEST FOR A UNIVARIATE STATISTICAL OUTLIER

A. Introduction

The object of this chapter is to present a point outlier theory from a different point of view than is usually given. The various methods described in Chapter II, the review of literature, are concerned solely with the identification and testing of an outlying observation as an end in itself. As we mentioned in Chapter I, Anscombe discussed the problem of subsequent estimation in a general way but did not give any specific results as to the possible bias and size of the mean square error of the estimate obtained subsequent to an outlier test.

In the present chapter, we are interested in what effect the rejection of an outlying observation might have on subsequent estimation and/or hypothesis testing. In particular, we are interested in

(i) the estimation of the population mean subsequent to a test for an outlying observation, and

(ii) the size and power of subsequent tests of hypotheses concerning the mean of the population.

We investigate these problems of estimation and hypothesis testing (a) when the scientist has performed the preliminary test for an outlying observation assuming no a priori informa-

tion, and (b) when the scientist has performed the preliminary test for an outlying observation assuming a priori information sufficient to identify a suspected outlier. In both situations we simplify the problem by considering only the case where one observation in the sample is suspect.

B. Estimation and Hypothesis Testing Subsequent to an Outlier Test Assuming No A Priori Information

1. Statement of the problem

Suppose that we have a random sample of N observations and wish to estimate the mean, μ_1, of the population and/or to test hypotheses about this population mean, say $H_0: \mu_1 = \mu_0$, subsequent to making a preliminary test for an outlying observation. We would like to use the statistic, $\tau = (x_N - \bar{x})/s$, whose exact distribution was found and tabled by Grubbs (1950), and which is referred to in this chapter as Grubbs' statistic, for this preliminary test for an outlier. However, the use of Grubbs' statistic leads to mathematical difficulties when we consider the subsequent problems of estimation and hypothesis testing. We, therefore, propose using a modified Thompson's (1935) criterion in place of Grubbs' criterion. It seems important to discuss in some detail the conditions under which the substitution of Thompson's statistic for Grubbs' statistic is possible.

As we mentioned in the review of literature, Pearson and

Sekar (1936) studied Thompson's criterion at great length and we now summarize briefly some of their findings. Letting $\tau^{(1)}$, $\tau^{(2)},\ldots,\tau^{(N)}$, represent the N values of τ in a sample arranged in descending order of magnitude taking account of signs, and using data based on 50 samples of size 5 drawn randomly from a normal population, they showed (a) that the total distribution of the 50 X 5 = 250 values of τ could be graduated by writing N = 5 in the equation $p(\tau)$ found by Thompson (1935), and (b) that the form of the total τ distribution at its extremes depended only on the distribution of $\tau^{(1)}$ and $\tau^{(N)}$, because, for some combinations of the sample size N and the percentage points, the algebraic upper limit for $\tau^{(2)}$ and the algebraic lower limit for $\tau^{(N-1)}$ do not extend into the tails of the total τ distribution. Thus for

$$\tau^{(1)} \geq \sqrt{(N-2)/2}$$

we may write for the probability law of $\tau^{(1)}$,

$$p(\tau^{(1)}) = N\, p(\tau),$$

and for

$$\tau^{(N)} < -\sqrt{(N-2)/2}$$

we may write

$$p(\tau^{(N)}) = N\, p(\tau).$$

Pearson and Sekar were then able to use Thompson's table to obtain the upper probability limits for $\tau^{(1)}$ and the lower probability limits for $\tau^{(N)}$, for some sample sizes.

Let us now investigate the combinations of percentage

points and sample sizes for which Thompson's criterion may be substituted for that of Grubbs. We would like to determine for which values of N,

$$\tau^{(1)}_{\alpha,N} \geq \sqrt{(N-2)/2}$$

when $\alpha = 0.01, 0.025, 0.05, 0.10$. We observe that both $\tau^{(1)}_{\alpha,N}$ and $\sqrt{(N-2)/2}$ are monotone in N. Using Table 1A given by Grubbs, we find that $\tau^{(1)}_{0.01,19} = 2.932$ and $\tau^{(1)}_{0.01,20} = 2.959$. Since $\sqrt{(19-2)/2} = 2.916$ and $\sqrt{(20-2)/2} = 3$, we conclude that the above inequality is true for $\alpha = 0.01$ if $N \leq 19$. In a similar fashion, we find that for $\alpha = 0.025$, $N \leq 16$; for $\alpha = 0.05$, $N \leq 15$; for $\alpha = 0.10$, $N \leq 11$.

The next question we ask ourselves is: can we construct an example where the inference drawn from Grubbs' procedure is not the same as the inference drawn from Thompson's procedure applied to $\tau^{(1)}$, irrespective of the validity of the condition $\tau^{(1)} \geq (N-2)/2$? Let us discuss this question for $\alpha = 0.05$ and $N = 15$. From Grubbs' Table 1A we have

$$\tau^{(1)}_{0.05,15} = 2.493$$

and from Thompson's table we have

$$\tau^{(1)}_{0.05,15} = 2.636.$$

The results of a significance test would differ provided the observed τ lies between 2.493 and 2.636. Hence we must construct an example where τ lies in this range for $N = 15$. Such an example can be obtained. Using Grubbs' table we have $P(\tau > 2.638) = 0.025$ and $P(\tau > 2.493) = 0.05$. Hence,

$P(2.493 < \tau < 2.636) = 0.025.$

Thus, for $\alpha = 0.05$ and $N = 15$, we can make the following statement. Roughly speaking, in about 25 samples out of 1000 drawn, the indiscriminate substitution of Thompson's procedure for testing the significance of $\tau^{(1)}$ would lead to a different result from the one that would be obtained using the exact procedure due to Grubbs.

With this background, we now proceed as follows. We assume that the observations $x_1, x_2, \ldots, x_N$ minus the arbitrary observation x_i, constitute a random sample of size N-1 from a normal population with mean μ_1 and variance σ^2, i.e. $N(\mu_1, \sigma^2)$, and that x_i is a random sample of size 1 from $N(\mu_2, \sigma^2)$. In other words, we assume that the normal universes have a common variance, σ^2, and universe means μ_1 and μ_2 which may or may not be equal. We are then in a situation which calls for an incompletely specified model in the sense of Bancroft (1964). If the universe means are, in fact, unequal, then x_i belongs to a universe different from that generating the other N-1 observations and is termed an outlier. Therefore, we first wish to test the hypothesis $H_1: \mu_1 = \mu_2$ versus $H_2: \mu_1 \neq \mu_2$. This test is accomplished by using Thompson's criterion in place of Grubbs' under the conditions specified above. For convenience, in the following section we repeat Thompson's criterion already discussed in the review of literature.

2. Thompson's criterion

Let $x_1, x_2, \ldots, x_N$ be a sample from a one-dimensional normal distribution with unknown parameters. Let x_i be an arbitrary one of the observations, $\bar{x} = \sum_{i=1}^{N} x_i/N$ and $s^2 = \sum_{i=1}^{N} (x_i-\bar{x})^2/N$. Thompson obtained the distribution of

$$\tau_i = (x_i-\bar{x})/s$$

and proposed that for a given α, values of x_i for which $|\tau| > \tau_\alpha$ be rejected, i.e. considered outliers, where τ_α is chosen so that for any i, $P(|\tau| > \tau_\alpha) = \alpha$. He determined τ_α for $\alpha = \phi/N$, where $\phi = 0.05, 0.10, 0.20$ and $N = 3(1)22, 32, 42, 102, 202, 1002$.

3. Estimation of μ_1 after Thompson's test for an outlying observation

If H_1 be rejected, we use $\bar{x}_{N-1} = (N\bar{x}-x_i)/(N-1)$ as the estimate of μ_1. If H_1 is not rejected, we use $\bar{x}_N = [(N-1)\bar{x}_{N-1}+x_i]/N$ as the estimate of μ_1. In estimation based on such an incompletely specified model, we are interested in the bias and the mean square error of the estimate, $\bar{x}^*$, where $\bar{x}^*$ is either $\bar{x}_{N-1}$ or $\bar{x}_N$, depending on the outcome of the preliminary test, which in our problem is Thompson's test in place of Grubbs' test.

4. Rule of procedure

In order to simplify subsequent mathematical manipulations, we use an amended form of Thompson's criterion for the

preliminary test for an outlier. Thompson showed that his criterion is related to Student's t test as follows:

$$t = \frac{\tau\sqrt{N-2}}{\sqrt{N-1-\tau^2}} \tag{4}$$

where $\tau = (x_i-\bar{x})/s$, $\bar{x} = \sum_{i=1}^{N} x_i/N$, $s^2 = \sum_{i=1}^{N}(x_i-\bar{x})^2/N$ and t is distributed with N-2 degrees of freedom. It can easily be shown that (4) becomes

$$t = \frac{\sqrt{N-1}\,(x_i-\bar{x}_{N-1})}{\sqrt{N}\,s_1} \tag{5}$$

where $s_1^2 = \sum_{i=1}^{N-1}(x_i-\bar{x}_{N-1})^2/(N-2)$; $\bar{x}_{N-1}$ being the mean of the N-1 observations that remain when the arbitrary observation x_i is removed from the sample, and the summation $\sum_{i=1}^{N-1}$ indicates that the term following is summed over the pruned sample. We also note that

$$P(|\tau| > \tau_\alpha) = \phi/N = \alpha = P(|t| > t_\alpha),$$

where t is Student's distribution with N-2 degrees of freedom, is related to τ by (4) and $t_\alpha = \tau_\alpha\sqrt{N-2}/\sqrt{N-1-\tau_\alpha^2}$. The rule of procedure is to calculate

$$t = \frac{\sqrt{N-1}\,(x_i-\bar{x}_{N-1})}{\sqrt{N}\,s_1}.$$

If t is non-significant at some preassigned significance level, say α, we use $\bar{x}_N$ as the estimate of μ_1. If t is significant, we use $\bar{x}_{N-1}$ as the estimate of μ_1, and conclude that x_i is an outlier.

5. Derivation of $E(\bar{x}^*)$

We know that

$$E(\bar{x}^*) = E(\bar{x}_{N-1} \mid |t| \geq t_\alpha)\ P(|t| \geq t_\alpha) + E(\bar{x}_N \mid |t| < t_\alpha)\ P(|t| < t_\alpha). \tag{6}$$

First we wish to find $E(\bar{x}_{N-1})$ if $|t| \geq t_\alpha$, i.e.

$$E\left(\frac{N\bar{x} - x_i}{N-1}\right), \text{ if } |t| \geq t_\alpha,$$

where t_α is the value on the t distribution corresponding to some preassigned significance level, say α. Let

$$y_1 = \frac{\bar{x}_{N-1} - \mu_1}{\sigma/\sqrt{N-1}},$$

$$y_2 = \frac{x_i - \mu_2}{\sigma},$$

and

$$W = \frac{(N-2)\ s_1^2}{\sigma^2}.$$

Then the joint distribution of y_1, y_2, W is

$$\frac{1}{2^{\frac{N}{2}} \pi\, \Gamma\left(\frac{N-2}{2}\right)}\ e^{-\frac{y_1^2 + y_2^2}{2}}\ e^{-\frac{W}{2}}\ W^{\frac{N-4}{2}}. \tag{7}$$

Let

$$v = \sqrt{\frac{N-1}{N}}\left(y_1 + \frac{y_2}{\sqrt{N-1}}\right),$$

$$u = \sqrt{\frac{N-1}{N}}\left(y_2 - \frac{y_1}{\sqrt{N-1}}\right),$$

$$W = W.$$

Then

$$\begin{aligned} y_1 &= \sqrt{\frac{N-1}{N}}\left(v - \frac{u}{\sqrt{N-1}}\right), \\ y_2 &= \sqrt{\frac{N-1}{N}}\left(u + \frac{v}{\sqrt{N-1}}\right), \end{aligned} \tag{8}$$

and the Jacobian of the transformation is 1. From (4) and (8) we obtain the joint distribution of u,v,W:

$$\frac{1}{2^{\frac{N}{2}}\pi\Gamma\left(\frac{N-2}{2}\right)} e^{-\frac{u^2 + v^2}{2}} e^{-\frac{W}{2}} W^{\frac{N-4}{2}}, \tag{9}$$

where $-\infty < u < \infty$, $W > 0$.

We can write

$$\bar{x}_{N-1} = Au + Bv + C$$

and

$$\frac{\sqrt{N-1}\ (x_i - \bar{x}_{N-1})}{\sqrt{N}\ s_1} = \frac{Du + E}{\sqrt{W}},$$

where

$A = -\sigma/\sqrt{N(N-1)}$, $B = \sigma/\sqrt{N}$, $C = \mu_1$, $D = \sqrt{N-2}$ and

$$E = \frac{\sqrt{(N-1)(N-2)}\ (\mu_2 - \mu_1)}{\sqrt{N}\ \sigma}.$$

Then $E(\bar{x}_{N-1})$ if $|t| \geq t_\alpha$ is given by

$$\frac{K}{P\left\{\left|\frac{Du + E}{\sqrt{W}}\right| \geq t_\alpha\right\}} \iiint (Au + Bv + C)\, e^{-\frac{u^2 + v^2}{2}}\, e^{-\frac{W}{2}}\, W^{\frac{N-4}{2}},$$

dudvdW,

where

$$K = \frac{1}{2^{\frac{N}{2}}\pi\Gamma\left(\frac{N-2}{2}\right)}$$

and the region of integration is given by

$$\left|\frac{Du + E}{\sqrt{W}}\right| \geq t_\alpha, \quad -\infty < v < \infty.$$

Integrating out the v, we have

$$\frac{K\sqrt{2\pi}}{P\left\{\left|\frac{Du + E}{\sqrt{W}}\right| \geq t_\alpha\right\}} \int_R \int (Au + C)\, e^{-\frac{u^2}{2}}\, e^{-\frac{W}{2}}\, W^{\frac{N-4}{2}}\, dudW, \quad (10)$$

where R, the region of integration, is given by

$$\left|\frac{Du + E}{\sqrt{W}}\right| \geq t_\alpha, \quad -\infty < u < \infty.$$

We now wish to find $E(\bar{x}_N)$ when $|t| < t_\alpha$. We note that

$$\bar{x}_N = Bv + F,$$

where

$$F = [(N-1)\mu_1 + \mu_2]/N \text{ and } B = \sigma/\sqrt{N}.$$

The expected value of $\bar{x}_N$ when $|t| < t_\alpha$ is given by

$$\frac{K}{P\left\{\left|\frac{Du + E}{\sqrt{W}}\right| < t_\alpha\right\}} \iiint (Bv + F)\, e^{-\frac{u^2 + v^2}{2}}\, e^{-\frac{W}{2}}\, W^{\frac{N-4}{2}}\, dudvdW.$$

Integrating out the v, we obtain

$$\frac{\sqrt{2\pi}\ K\ F}{P\left\{\left|\frac{Du + E}{\sqrt{W}}\right| < t_\alpha\right\}} \int\int_{R'} e^{-\frac{u^2}{2}}\ e^{-\frac{W}{2}}\ W^{\frac{N-4}{2}}\ dudW, \tag{11}$$

where the region of integration, R', is given by

$$\left|\frac{Du + E}{\sqrt{W}}\right| < t_\alpha,\ -\infty < u < +\infty.$$

Combining (6), (10) and (11), we have

$$E(\bar{x}^*) = \sqrt{2\pi}K\left[\int\int_R (Au + C)\ e^{-\frac{u^2}{2}}\ e^{-\frac{W}{2}}\ W^{\frac{N-4}{2}}\ dudW\right.$$

$$\left. + F\int\int_{R'} e^{-\frac{u^2}{2}}\ e^{-\frac{W}{2}}\ W^{\frac{N-4}{2}}\ dudW\right], \tag{12}$$

where the regions of integration, R and R', may be replaced by the limits

$$\begin{bmatrix} -\infty < u < +\infty \\ 0 < W < \dfrac{(Du + E)^2}{t_\alpha^2} = Q \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -\infty < u < +\infty \\ W > Q \end{bmatrix}$$

respectively.

As a partial check on this result we let $t_\alpha = 0$, i.e. we always reject the hypothesis, and we find $E(\bar{x}^*) = \mu_1$. If we take the limit as $t_\alpha \longrightarrow +\infty$, i.e. we never reject the hypothesis, then $E(\bar{x}^*) \longrightarrow [(N-1)\mu_1 + \mu_2]/N$.

To simplify (12) we use repeated integration by parts and integrate out the W. We obtain

$$E(\bar{x}^*) = \frac{1}{\sqrt{2\pi}}\left\{F\int_{-\infty}^{\infty} e^{-\frac{u^2}{2}}\ e^{-\frac{Q}{2}}\ \sum_{i=0}^{\frac{N-4}{2}}\left(\frac{Q}{2}\right)^i \frac{1}{i!}\,du\right.$$

$$\left. + \int_{-\infty}^{\infty} (Au + C)\, e^{-\frac{u^2}{2}}\left[1 - e^{-\frac{Q}{2}}\ \sum_{i=0}^{\frac{N-4}{2}}\left(\frac{Q}{2}\right)^i \frac{1}{i!}\right] du\right\},$$

or

$$E(\bar{x}^*) = \mu_1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(Au + C - F)\, e^{-\frac{u^2 + Q}{2}}\ \sum_{i=0}^{\frac{N-4}{2}}\left(\frac{Q}{2}\right)^i \frac{1}{i!}\ du,$$

for N an even integer ≥ 4.

To carry out the integration, let $Z = u + \lambda$, where $\lambda = E/D$. Then $u = Z - \lambda$, $Q = (N-2)(u + \lambda)^2/t_\alpha^2$ and $E(\bar{x}^*)$ becomes

$$\mu_1 - \frac{M}{\sqrt{2\pi}}\sum_{i=0}^{\frac{N-4}{2}}\left(\frac{D^2}{2t_\alpha^2}\right)^i \frac{1}{i!}\int_{-\infty}^{\infty} Z^{2i}\,(AZ - A\lambda + C - F)$$

$$e^{-\frac{(Z - G)^2}{2H}}\,dZ, \qquad (13)$$

where

$$M = e^{-\frac{\lambda^2 D^2}{2(t_\alpha^2 + D^2)}},\quad H = \frac{t_\alpha^2}{t_\alpha^2 + D^2} \text{ and } G = \frac{\lambda t_\alpha^2}{t_\alpha^2 + D^2}.$$

However,

$$\int_{-\infty}^{\infty} Z^{2i+1}\ e^{-\frac{(Z - G)^2}{2H}}\,dZ = \sqrt{2\pi H}\ \mu'_{2i+1}$$

and

$$\int_{-\infty}^{\infty} Z^{2i} \; e^{-\frac{(Z - G)^2}{2H}} \, dZ = \sqrt{2\pi H}\, \mu'_{2i},$$

where μ'_{2i} and μ'_{2i+1} are the (2i)th and (2i+1)th moments about the origin of a normal distribution with mean G and variance H. Then (13) becomes

$$E(\bar{x}^*) = \mu_1 - \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{N-1}{2t_\alpha^2}\right)^i \frac{1}{i!}\left[A\mu'_{2i+1} - A\lambda\mu'_{2i} + C\mu'_{2i} - F\mu'_{2i}\right] \frac{t_\alpha e^{-\frac{\lambda^2(N-1)}{2(t_\alpha^2 + N - 1)}}}{\sqrt{t_\alpha^2 + N - 1}}, \qquad (14)$$

for N an even integer ≥ 4.

The bias in $\bar{x}^*$ as an estimator of μ_1 is given by

$$b = -\sum_{i=0}^{\frac{N-4}{2}} \left(\frac{N-1}{2t_\alpha^2}\right)^i \frac{1}{i!}\left[A\mu'_{2i+1} - A\lambda\mu'_{2i} + C\mu'_{2i} - F\mu'_{2i}\right] \frac{t_\alpha e^{-\frac{\lambda^2(N-1)}{2(t_\alpha^2 + N - 1)}}}{\sqrt{t_\alpha^2 + N - 1}}, \qquad (15)$$

for N an even integer ≥ 4.

As a partial check we let $t_\alpha = 0$, then $b = 0$. If we let $t_\alpha \longrightarrow +\infty$, $b \longrightarrow (\mu_2 - \mu_1)/N$.

It was verified computationally that the bias depends only on δ where

$$\delta = \frac{|\mu_2 - \mu_1|}{\sigma},$$

by varying the means and standard deviation but keeping δ constant. In Table 2 bias values are given for δ = 0.5, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0; α = 0.01, 0.5; N = 6, 8, 10, 12, 14, 16, 18, 20, 22, 24. However, substitution of Thompson's procedure for Grubbs' procedure is strictly valid up to the values of N underlined in the table.

Table 2. Bias, b, of estimator based on preliminary test

		δ						
	N	0.5	1.0	1.5	2.0	3.0	4.0	5.0
α = 0.01								
	6	0.181	0.159	0.230	0.289	0.356	0.341	0.262
	8	0.062	0.121	0.172	0.210	0.229	0.174	0.093
	10	0.050	0.098	0.138	0.165	0.165	0.106	0.043
	12	0.042	0.082	0.115	0.136	0.127	0.071	0.024
	14	0.037	0.071	0.099	0.115	0.103	0.053	0.015
	16	0.032	0.062	0.086	0.100	0.086	0.041	0.011
	18	0.029	0.056	0.077	0.088	0.074	0.033	0.008
	20	0.026	0.050	0.069	0.079	0.064	0.027	0.006
	22	0.024	0.046	0.063	0.071	0.057	0.023	0.005
	24	0.022	0.042	0.058	0.065	0.051	0.020	0.004
α = 0.05								
	6	0.068	0.127	0.169	0.187	0.158	0.087	0.032
	8	0.052	0.095	0.122	0.129	0.091	0.038	0.009
	10	0.042	0.076	0.096	0.098	0.062	0.021	0.004
	12	0.035	0.063	0.079	0.079	0.047	0.014	0.002
	14	0.030	0.054	0.067	0.066	0.037	0.010	0.001
	16	0.027	0.047	0.058	0.057	0.030	0.008	0.001
	18	0.024	0.042	0.051	0.049	0.026	0.006	0.001
	20	0.021	0.038	0.046	0.044	0.022	0.005	0.001
	22	0.019	0.034	0.042	0.040	0.020	0.004	0.000
	24	0.018	0.032	0.038	0.036	0.017	0.003	0.000

6. Mean square error of $\bar{x}^*$

The notation introduced previously will be used in this section. We know that

$$MSE(\bar{x}^*) = E(\bar{x}^{*2}) - E^2(\bar{x}^*) + (Bias)^2.$$

Therefore, the only new calculation involved in evaluating the mean square error of $\bar{x}^*$ is to compute $E(\bar{x}^{*2})$ and subtract from it the terms of $E^2(\bar{x}^*)$ which are not contained in $(Bias)^2$. To find $E(\bar{x}^{*2})$, we write

$$\begin{aligned} E(\bar{x}^{*2}) &= E(\bar{x}_{N-1}^2 \,|\, |t| \geq t_\alpha)\, P(|t| \geq t_\alpha) \\ &\quad + E(\bar{x}_N^2 \,|\, |t| < t_\alpha)\, P(|t| < t_\alpha) \\ &= E\left[(Au + Bv + C)^2 \,\middle|\, \left|\frac{Du + E}{\sqrt{W}}\right| \geq t_\alpha\right] P\left[\left|\frac{Du + E}{\sqrt{W}}\right| \geq t_\alpha\right] \\ &\quad + E\left[(Bv + F)^2 \,\middle|\, \left|\frac{Du + E}{\sqrt{W}}\right| < t_\alpha\right] P\left[\left|\frac{Du + E}{\sqrt{W}}\right| < t_\alpha\right], \end{aligned}$$

or

$$\begin{aligned} E(\bar{x}^{*2}) &= \int_0^Q \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (Au + Bv + C)^2 \, f(u,v,W)\, du\,dv\,dW \\ &\quad + \int_0^Q \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (Bv + F)^2 \, f(u,v,W)\, du\,dv\,dW, \end{aligned}$$

where $f(u,v,W)$ is given by (9).

Integrating out the v and then the W, we have

$$E(\bar{x}^{*2}) = \frac{1}{\sqrt{2\pi}} \left\{ \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} \left[[(Au + C)^2 + B^2] - e^{-\frac{Q}{2}} [(Au + C)^2 - F^2] \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{Q}{2}\right)^i \frac{1}{i!} \right] du \right\}. \quad (16)$$

Subtracting from $E(\bar{x}^{*2})$ those terms of $E^2(\bar{x}^*)$ which are not in $(\text{Bias})^2$, we have

$$MSE(\bar{x}^*) = A^2 + B^2 + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (F^2 - 2CF + C^2 - A^2u^2) \, e^{-\frac{u^2 + Q}{2}} \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{Q}{2}\right)^i \frac{1}{i!} \, du.$$

Integrating out the u by making the substitution $Z = u + \lambda$, and proceeding as in Part 5, we have

$$MSE(\bar{x}^*) = A^2 + B^2 + \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{N-1}{2t_\alpha^2}\right)^i \frac{1}{i!} [(F - C)^2 \, \mu'_{2i} - A^2(\mu'_{2i+2} - 2\lambda\mu'_{2i+1} + \lambda^2\mu'_{2i})] \frac{t_\alpha e^{-\frac{\lambda^2(N-1)}{2(t_\alpha^2 + N - 1)}}}{\sqrt{t_\alpha^2 + N - 1}}. \tag{17}$$

When $t_\alpha \longrightarrow +\infty$, $MSE(\bar{x}^*) \longrightarrow \sigma^2/N + (\mu_1 - \mu_2)^2/N^2$; when $t_\alpha = 0$, $MSE(\bar{x}^*) = \sigma^2/(N-1)$. This serves as a partial check for (17).

Values of $MSE(\bar{x}^*)$ are given in Table 3 for δ = 0.5, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0; α = 0.01, 0.05; N = 6, 8, 10, 12, 14, 16, 18, 20, 22, 24. The value of δ is defined on page 43. Again, the substitution of Thompson's procedure for Grubbs' procedure is strictly valid only up to the values of N underlined in the table.

Table 3. Mean square error of estimator based on preliminary test

		δ						
		0.5	1.0	1.5	2.0	3.0	4.0	5.0
	N							
α = 0.01								
	6	0.174	0.196	0.229	0.273	0.367	0.430	0.431
	8	0.129	0.142	0.161	0.185	0.229	0.238	0.210
	10	0.103	0.111	0.123	0.139	0.163	0.160	0.137
	12	0.085	0.091	0.100	0.111	0.126	0.120	0.104
	14	0.073	0.077	0.084	0.092	0.102	0.096	0.084
	16	0.063	0.067	0.072	0.078	0.085	0.080	0.071
	18	0.056	0.059	0.063	0.068	0.073	0.068	0.062
	20	0.051	0.053	0.056	0.060	0.064	0.060	0.055
	22	0.046	0.048	0.051	0.054	0.057	0.053	0.049
	24	0.042	0.044	0.046	0.049	0.051	0.048	0.045
α = 0.05								
	6	0.179	0.200	0.228	0.257	0.287	0.269	0.233
	8	0.132	0.144	0.160	0.175	0.184	0.167	0.150
	10	0.104	0.112	0.123	0.132	0.135	0.123	0.113
	12	0.086	0.092	0.099	0.105	0.106	0.097	0.092
	14	0.074	0.078	0.083	0.087	0.087	0.081	0.078
	16	0.064	0.067	0.071	0.075	0.074	0.070	0.067
	18	0.057	0.059	0.062	0.065	0.065	0.061	0.059
	20	0.051	0.053	0.056	0.058	0.057	0.054	0.052
	22	0.046	0.048	0.050	0.052	0.051	0.049	0.048
	24	0.042	0.044	0.046	0.047	0.046	0.044	0.043

7. Test for H_0: $\mu_1 = \mu_0$ after Thompson's test for an outlying observation has been made

We now give a test procedure to be used to test a hypothesis about the mean of a normal population subsequent to a test for an outlying observation. We again use the amended form of Thompson's criterion in place of Grubbs' criterion for the preliminary test for an outlier. We propose the following test procedure:

(i) If $|t| \geq t_\alpha$, test the hypothesis H_0: $\mu_1 = \mu_0$ versus H_a: $\mu_1 \neq \mu_0$ by applying the t test to the statistic

$$t_1 = \frac{\bar{x}_{N-1} - \mu_0}{s_1/\sqrt{N-1}}$$

with N-2 degrees of freedom and significance level α_1.

(ii) If $|t| < t_\alpha$, then assuming $\mu_1 = \mu_2 = \mu_{12}$, say, test the hypothesis H_0: $\mu_{12} = \mu_0$ versus H_a: $\mu_{12} \neq \mu_0$ by applying the t test to the statistic

$$t_2 = \frac{\bar{x}_{12} - \mu_0}{s_1/\sqrt{N}}$$

with N-2 degrees of freedom and significance level α_2, where $\bar{x}_{12} = \bar{x} = [(N-1)\bar{x}_{N-1} + x_i]/N$.

8. Power of the test procedure given in Part 7

The power of a test, against a particular alternative, is the probability that the test will reject the null hypothesis if the alternative is true. It is also defined as one minus the probability of committing a type II error where a type II error is the error committed by accepting a false hypothesis. We obtain the power P as the sum of two mutually exclusive components corresponding to the mutually exclusive alternatives given in Part 7. That is,

$$P = \Pr(|t| \geq t_\alpha \text{ and } |t_1| \geq \phi_1) + \Pr(|t| < t_\alpha \text{ and } |t_2| \geq \phi_2), \tag{18}$$

where t_α, ϕ_1, and ϕ_2 are critical values of the t distribution

corresponding to significance levels α, α_1, and α_2 respectively.

To evaluate the first term on the right hand side of (18), we start with the joint distribution of u,v,W given by (9). Then let

$$T = \frac{\sqrt{N-1}\,(x_i - \bar{x}_{N-1})}{\sqrt{N}\,s_1} = \frac{au + b}{\sqrt{W}},$$

$$T_1 = \frac{\bar{x}_{N-1} - \mu_0}{s\,/\sqrt{N-1}} \quad \frac{cv - du + e}{\sqrt{W}},$$

where

$$a = \sqrt{N-2},\ b = \sqrt{(N-1)(N-2)}\,(\mu_2 - \mu_1)/\sqrt{N}\,\sigma,$$

$$c = \sqrt{(N-1)(N-2)/N},\ d = \sqrt{(N-2)/N},\ e = \sqrt{(N-1)(N-2)}(\mu_1-\mu_0)/\sigma.$$

Then

$$u = \frac{\sqrt{W}\,T - b}{a},$$

$$v = \frac{a\sqrt{W}\,T_1 - ae + d\sqrt{W}\,T - bd}{ac},$$

$$W = W,$$

and the Jacobian of the transformation is W/ac. We now have

$$f(T,T_1,W) = K^* W^{\frac{N-2}{2}} e^{-\frac{fW - 2gW^{\frac{1}{2}} + h}{2}}, \qquad (19)$$

where

$$K^* = \frac{\sqrt{N}}{2^{\frac{N}{2}}\,\pi\,\sqrt{(N-1)}\,(N-2)\,\Gamma\left(\frac{N-2}{2}\right)},$$

$$f = \frac{N}{(N-1)(N-2)}[T^2 + \frac{2}{\sqrt{N}} T_1 T + T_1^2 + \frac{(N-1)(N-2)}{N}],$$

$$g = \frac{1}{\sigma\sqrt{(N-1)(N-2)}} [(\mu_2 - \mu_0)\sqrt{N}\, T + (N\mu_1 - \mu_1 - N\mu_0 + \mu_2)T_1],$$

$$h = \frac{1}{\sigma^2} [(\mu_2 - \mu_1)^2 + N(\mu_1 - \mu_0)^2 + 2(\mu_1 - \mu_0)(\mu_2 - \mu_1)].$$

Completing the square in the exponent, we have

$$f(T,T_1,W) = K^* \; e^{-\frac{1}{2}(h - \frac{g^2}{f})} \; W^{\frac{N-2}{2}} \; e^{-\frac{f}{2}(W^{\frac{1}{2}} - \frac{g}{f})^2}. \qquad (20)$$

To integrate out the W, we note that

$$\int_0^\infty W^{\frac{N-2}{2}} \; e^{-\frac{f}{2}(W^{\frac{1}{2}} - \frac{g}{f})^2} \, dW = \frac{2^{\frac{N+2}{2}}}{f^{\frac{N}{2}}} \sum_{i=0}^{N-1} \binom{N-1}{i} \left(\frac{g}{\sqrt{2f}}\right)^{N-1-i} \int_{-g/\sqrt{2f}}^\infty y^i \; e^{-y^2} \, dy,$$

where

$$y = \sqrt{\frac{f}{2}} \, (W^{\frac{1}{2}} - \frac{g}{f})$$

and

$$\int_{-g/\sqrt{2f}}^\infty y^i \; e^{-y^2} \, dy = (1-s) \, \frac{(1;2;r)\sqrt{\pi}}{2^{r+1}} \left[1 - \phi\left(\frac{-g}{\sqrt{2f}}\right)\right] + e^{-\frac{g^2}{2f}} \sum_{\nu=0}^{r-1} \frac{(i-1;-2;\nu)}{2^{\nu+1}} \left(\frac{-g}{\sqrt{2f}}\right)^{i-2\nu-1},$$

where

$$i = 2r - s,\ s = 0 \text{ or } 1,\ \Phi\left(\frac{-g}{\sqrt{2f}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{-g/\sqrt{2f}} e^{-t^2}\, dt,$$

$$(m;d;\nu) = m(m+d)(m+2d) \ldots (m+d[\nu-1]),\ (m;d;0) = 1,$$

$$(m;-d;\nu) = d^{\nu}\, \Gamma\left(\frac{m}{d}+1\right) / \Gamma\left(\frac{m}{d}+1-\nu\right).$$

Then

$$f(T,T_1) = \frac{2\sqrt{N}}{f^{\frac{N}{2}} \pi\ (N-2)\sqrt{N-1}\,\Gamma\left(\frac{N-2}{2}\right)}\ e^{-\frac{1}{2}\left(h - \frac{g^2}{f}\right)} \sum_{i=0}^{N-1} \binom{N-1}{i}$$

$$\left(\frac{g}{\sqrt{2f}}\right)^{N-1-i} \left\{(1-s)\ \frac{(1;2;r)\sqrt{\pi}}{2^{r+1}} \left[1 - \Phi\left(\frac{-g}{\sqrt{2f}}\right)\right]\right.$$

$$\left. + e^{-\frac{g^2}{2f}} \sum_{\nu=0}^{r-1} \frac{(i-1;-2;\nu)}{2^{\nu+1}} \left(\frac{-g}{\sqrt{2f}}\right)^{i-2\nu-1}\right\}.$$

The formula for the first power component, P_1, is given by

$$P_1 = \int_{|T| \geq t_\alpha} \int_{|T_1| \geq \phi_1} f(T,T_1)\ dT dT_1.$$

To obtain the second component of the power, P_2, we need the joint distribution of T and T_2 where

$$T = \frac{\sqrt{N-1}\ (x_i - \bar{x}_{N-1})}{\sqrt{N-1}\ s_1} = \frac{au + b}{\sqrt{W}},$$

$$T_2 = \frac{\bar{x}_{12} - \mu_0}{s_1/\sqrt{N}} = \frac{av + k}{\sqrt{W}},$$

where

$$k = \frac{\sqrt{N-2}}{\sqrt{N}\,\sigma}\,(N\mu_1 - \mu_1 + \mu_2 - N\mu_0).$$

We start with the joint distribution of u,v,W given by (9), and make the transformation indicated above. We obtain

$$f(T,T_2,W) = k^* \; W^{\frac{N-2}{2}} \; e^{-\frac{pW - 2qW^{\frac{1}{2}} + r}{2}},$$

where

$$k^* = [2^{\frac{N}{2}}\pi\,(N-2)\,\Gamma\left(\frac{N-2}{2}\right)]^{-1},$$

$$p = \frac{1}{N-2}\,[N - 2 + T^2 + T_2^2],$$

$$q = \frac{1}{\sqrt{N(N-2)}\sigma}\,[\sqrt{N-1}(\mu_2 - \mu_1)T + (N\mu_1 - \mu_1 + \mu_2 - N\mu_0)T_2],$$

$$r = \frac{1}{\sigma^2}\,[(\mu_2 - \mu_1)^2 + N(\mu_1 - \mu_0)^2 + 2(\mu_1 - \mu_0)(\mu_2 - \mu_1)].$$

Completing the square in the exponent, we obtain

$$f(T,T_2,W) = k^* \; e^{-\frac{1}{2}(r - \frac{q^2}{p})} \; W^{\frac{N-2}{2}} \; e^{-\frac{p}{2}(W^{\frac{1}{2}} - \frac{q}{p})^2}.$$

Integrating out the W, we have

$$f(T,T_2) = \frac{2e^{-\frac{1}{2}(r - \frac{q^2}{p})}}{p^{\frac{N}{2}}\pi\,(N-2)\,\Gamma\left(\frac{N-2}{2}\right)} \sum_{i=0}^{N-1} \binom{N-1}{i} \left(\frac{q}{\sqrt{2p}}\right)^{N-1-i} \left\{(1-s')\right.$$

$$\left. \cdot \frac{(1;2;r')\sqrt{\pi}}{2^{r'+1}} \left[1 - \Phi\left(\frac{-q}{\sqrt{2p}}\right)\right] + e^{-\frac{q^2}{2p}}\right.$$

$$\cdot \sum_{\nu=0}^{r'-1} \frac{(i-1;-2;\nu)}{2^{\nu+1}} \left(\frac{-q}{\sqrt{2p}}\right)^{i-1-2\nu} \Bigg\},$$

where

$$i = 2r' - s',\ s' = 0 \text{ or } 1,\ \phi\left(\frac{-q}{\sqrt{2p}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{-q/\sqrt{2p}} e^{-t^2}\, dt.$$

Then the formula for the second power component, P_2, is given by

$$P_2 = \int_{|T| < t_\alpha} \int_{|T_2| \geq \phi_2} f(T,T_2)\, dT dT_2.$$

The power defined by $P = P_1 + P_2$ can now be written as

$$P = \int_{|T| \geq t_\alpha} \int_{|T_1| \geq \phi_1} f(T,T_1)\, dT dT_1 + \int_{|T| < t_\alpha} \int_{|T_2| \geq \phi_2} f(T,T_2)\, dT dT_2,$$

where $f(T,T_1)$ and $f(T,T_2)$ are given above.

Integration difficulties prevent an explicit evaluation of the power, that is, we cannot represent the power in closed form using well-known functions. Numerical evaluations are not feasible unless one makes use of an electronic computer. A program is now being prepared by the Iowa State University Computer Center which will give some numerical values for the power.

C. Estimation and Hypothesis Testing Subsequent to an Outlier Test Assuming A Priori Information Sufficient to Identify a Suspected Outlier

1. Statement of the problem

A simpler problem than the more general one described in Section B would be the case where one already has a priori information of a kind that makes it possible to say in advance that any observation in a sample to be drawn would be suspect if it were greater than C (or less than C), where C is known from a priori information. That is, we envisage a situation where the scientist might very well know in advance the limit or optimum size of the observations and would suspect that any observation greater (or less) than this limit might be from a normal population different from that generating the remaining observations. This problem is easier to solve because the preliminary test for an outlier is Student's t test when the variance is unknown and the Z test, based on normal theory, when the variance is known. Again, as in Section B, we further simplify the problem by considering only the case where one observation in the sample is less than C. By definition, this observation is suspected of belonging to another population. A similar procedure could be used when one observation is greater than C.

Suppose we have a random sample of N observations and wish to estimate the mean, μ_1, of the population or to test

hypotheses about this population mean, say H_0: $\mu_1 = \mu_0$. From a priori information we suspect that the smallest observation, say x_N, is an outlier. To take into account this added uncertainty we assume that the observations, $x_1, x_2, \ldots, x_{N-1}$, constitute a random sample of size N-1 from a normal population with mean μ_1 and variance σ^2, i.e. $N(\mu_1, \sigma^2)$, and that x_N is a random sample of size 1 from $N(\mu_2, \sigma^2)$. In other words, we assume that the normal universes have a common variance, σ^2, and universe means, μ_1 and μ_2, which may or may not be equal. We are then in a situation which calls for an incompletely specified model in the sense of Bancroft (1964). If the universe means are, in fact, unequal, then x_N belongs to a different universe and is termed an outlier. Therefore, we first wish to test the hypothesis H_1: $\mu_1 = \mu_2$ versus H_2: $\mu_1 > \mu_2$. This test is accomplished by using one of the statistics defined below.

(i) If σ^2 is known, we use

$$Z_1 = \frac{\bar{x}_{N-1} - x_N}{\sigma\sqrt{1 + 1/(N-1)}},$$

where $\bar{x}_{N-1} = \sum_{i=1}^{N-1} x_i/(N-1)$. The value of Z_1 is calculated from the data and compared with $Z_{\alpha_1} = \phi$, where Z_{α_1} or ϕ is the critical value of the normal distribution with significance level α_1. If $Z_1 \geq \phi$ we reject the hypothesis H_1 and in view of this evidence conclude that x_N is indeed an outlier. If $Z_1 < \phi$ we conclude that we have no reason to believe that x_N is an outlier.

(ii) If σ^2 is unknown, we use

$$t_1 = \frac{\bar{x}_{N-1} - x_N}{s\sqrt{1 + 1/(N-1)}}$$

where $\bar{x}_{N-1} = \sum_{i=1}^{N-1} x_i/(N-1)$ and $s^2 = \sum_{i=1}^{N-1} (x_i - \bar{x}_{N-1})^2/(N-2)$. The calculated value of t_1 is compared with $t_{N-2,\alpha_1} = \phi$, where t_{N-2,α_1} or ϕ is the critical value of the t distribution with N-2 degrees of freedom and significance level α_1. If $t_1 \geq \phi$ we reject the hypothesis H_1 and conclude that x_N is an outlier. If $t_1 < \phi$ we conclude that we have have no reason to believe that x_N is an outlier.

2. Estimation of μ_1 after a preliminary test for an outlying observation

If H_1 is rejected, we use $\bar{x}_{N-1}$ as the estimate of μ_1. If H_1 is not rejected, we use $\bar{x}_N = [(N-1)\bar{x}_{N-1} + x_N]/N$ as the estimate of μ_1. In estimation based on such an incompletely specified model, we are interested in the bias and the mean square error of the estimate, $\bar{x}^*$, where $\bar{x}^*$ is either $\bar{x}_N$ or $\bar{x}_{N-1}$, depending on the outcome of the preliminary test.

3. Derivation of $E(\bar{x}^*)$ and the mean square error of $\bar{x}^*$ (σ^2 known)

We assume in this section that σ^2 is known and equal to one. First, we wish to find $E[(N-1)\bar{x}_{N-1} + x_N]/N$ if $Z_1 < \phi$, where ϕ is the critical value of the normal distribution corresponding to some preassigned significance level, say α_1.

The joint distribution of $\bar{x}_{N-1}$ and x_N is given by

$$\frac{\sqrt{N-1}}{2\pi} e^{-\frac{1}{2}[(N-1)(\bar{x}_{N-1} - \mu_1)^2 + (x_N - \mu_2)^2]}. \tag{21}$$

Let us make the transformation of variables

$$Z = \frac{(\bar{x}_{N-1} - x_N)\sqrt{N-1}}{\sqrt{N}},$$

$$V = \frac{(N-1)\bar{x}_{N-1} + x_N}{N}.$$

Then

$$\bar{x}_{N-1} = V + \frac{Z}{\sqrt{N(N-1)}},$$

$$x_N = V - \frac{\sqrt{N-1}\, Z}{\sqrt{N}},$$

and the joint distribution of Z and V is given by

$$\frac{\sqrt{N}}{2\pi} e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} e^{-\frac{N}{2}\left[V - \frac{(N-1)\mu_1 + \mu_2}{N}\right]^2}. \tag{22}$$

The expected value of V given $Z < \phi$ is given by

$$\frac{\int_{-\infty}^{\phi}\int_{-\infty}^{\infty} V\, f(V,Z)\, dV dZ}{P(Z < \phi)}. \tag{23}$$

Integrating out the V in (23), we have

$$E(V|Z < \phi) = \frac{(N-1)\mu_1 + \mu_2}{NP(Z < \phi)\sqrt{2\pi}} \int_{-\infty}^{\phi} e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ. \tag{24}$$

However,

$$P(Z < \phi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\phi} e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ. \qquad (25)$$

Therefore, we have

$$E(V|Z < \phi) = \frac{(N-1)\mu_1 + \mu_2}{N}. \qquad (26)$$

We now wish to find the expected value of $\bar{x}_{N-1}$ when $Z \geq \phi$. We again start with the joint distribution of $\bar{x}_{N-1}$ and x_N, and this time let

$$Z = \frac{(\bar{x}_{N-1} - x_N)\sqrt{N-1}}{\sqrt{N}},$$

$$W = \bar{x}_{N-1}.$$

Then the joint distribution of W and Z is

$$\frac{\sqrt{N}}{2\pi} e^{-\frac{N}{2}\left[W - \frac{(N-1)\mu_1 + \mu_2 + \sqrt{\frac{N}{N-1}}Z}{N}\right]^2} e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} \qquad (27)$$

and the expected value of W given $Z \geq \phi$ is

$$\frac{1}{P(Z \geq \phi)\sqrt{2\pi}\, N} \int_{\phi}^{\infty} \left[(N-1)\mu_1 + \mu_2 + \sqrt{\frac{N}{N-1}}\, Z\right] \cdot e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ.$$

However,

$$P(Z \geq \phi) = \frac{1}{\sqrt{2\pi}} \int_{\phi}^{\infty} e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ. \qquad (28)$$

Therefore,

$$E(W|Z \geq \phi) = \frac{1}{P(Z \geq \phi)}\left[\frac{[(N-1)\mu_1 + \mu_2]\, P(Z \geq \phi)}{N}\right.$$

$$\left. + \frac{1}{\sqrt{2N(N-1)\pi}} \int_{\phi}^{\infty} Z\, e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ\right]. \qquad (29)$$

We know that

$$E(\bar{x}^*) = E(V|Z < \phi)\, P(Z < \phi) + E(W|Z \geq \phi)\, P(Z \geq \phi). \qquad (30)$$

From equations (26), (29) and (30), we obtain

$$E(\bar{x}^*) = \frac{(N-1)\mu_1 + \mu_2}{N}$$

$$+ \frac{1}{\sqrt{2N(N-1)\pi}} \int_{\phi}^{\infty} Z\, e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ. \qquad (31)$$

As a partial check we note that if $\phi \longrightarrow -\infty$, $E(\bar{x}^*) \longrightarrow \mu_1$; if $\phi \longrightarrow +\infty$, $E(\bar{x}^*) \longrightarrow [(N-1)\mu_1 + \mu_2]/N$.

The expression for $E(\bar{x}^*)$ given by (31) may be simplified as follows. Let

$$U = Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}},$$

then (31) becomes

$$E(\bar{x}^*) = \frac{(N-1)\mu_1 + \mu_2}{N} + \frac{1}{\sqrt{2N(N-1)\pi}} \int (U + \frac{\sqrt{N-1}\ \delta}{\sqrt{N}})\ e^{-\frac{U^2}{2}}\ dU$$

where $\mu_1 - \mu_2 = \delta \geq 0$, and the limits of integration are given by $U + \frac{\sqrt{N-1}\delta}{\sqrt{N}} \geq \phi$. Integrating the term $U\ e^{-U^2/2}$ over these limits, the above expression becomes

$$E(\bar{x}^*) = \frac{(N-1)\mu_1 + \mu_2}{N} + \frac{1}{\sqrt{2N(N-1)\pi}}\ e^{-\frac{1}{2}(\phi - \frac{\sqrt{N-1}\delta}{\sqrt{N}})^2}$$

$$+ \frac{\delta}{\sqrt{2\pi}\ N} \int_{\phi - \frac{\sqrt{N-1}\delta}{\sqrt{N}}}^{\infty} e^{-\frac{U^2}{2}}\ dU. \qquad (32)$$

If the mean of the sampling distribution of a statistic equals the corresponding population parameter, the statistic is called an unbiased estimator of the parameter; otherwise it is said to be biased. The bias of $\bar{x}^*$ in estimating μ_1 is given by

$$b = -\frac{\delta}{N} + \frac{1}{\sqrt{2N(N-1)\pi}}\ e^{-\frac{1}{2}(\phi - \frac{\sqrt{N-1}\ \delta}{\sqrt{N}})^2}$$

$$+ \frac{\delta}{\sqrt{2\pi}\ N} \int_{\phi - \frac{\sqrt{N-1}\ \delta}{\sqrt{N}}}^{\infty} e^{-\frac{U^2}{2}}\ dU. \qquad (33)$$

When $\phi \longrightarrow -\infty$, $b \longrightarrow 0$ and the bias disappears. It is also evident that when a bias is present it diminishes as N in-

creases. As a further check, we note that when $\phi \longrightarrow +\infty$ the bias is given by $(\mu_2 - \mu_1)/N$. Numerical values for the bias function when $N = 6$, $\delta = 0, 1, 3$ and $\phi = 0, 2, 3, \infty$ are given in Table 4.

Table 4. Bias values

ϕ	δ = 0	δ = 1	δ = 3
0	0.073	0.019	0.016
2	0.010	-0.102	-0.103
3	0.001	-0.490	-0.215
∞	0.000	-0.167	-0.500

To evaluate the mean square error of $\bar{x}^*$ we need to derive $E(\bar{x}^{*2})$. We know that

$$E(\bar{x}^{*2}) = E(\bar{x}^{*2} | Z < \phi)\, P(Z < \phi) + E(\bar{x}^{*2} | Z \geq \phi)\, P(Z \geq \phi),$$

or in the notation used previously,

$$E(\bar{x}^{*2}) = E(V^2 | Z < \phi)\, P(Z < \phi) + E(W^2 | Z \geq \phi)\, P(Z \geq \phi). \tag{34}$$

Using (22) we have

$$E(V^2 | Z < \phi) P(Z < \phi) = \frac{\sqrt{N}}{2\pi} \int_{-\infty}^{\phi} \int_{-\infty}^{\infty} V^2 \, e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2}$$

$$\cdot e^{-\frac{N}{2}\left[V - \frac{(N-1)\mu_1 + \mu_2}{N}\right]^2} dVdZ.$$

Integrating out the V, we obtain

$$E(V^2 | Z < \phi)P(Z < \phi) = \frac{N + [(N-1)\mu_1 + \mu_2]^2}{N^2} P(Z < \phi). \qquad (35)$$

Similarly, using (27) we have

$$E(W^2 | Z \geq \phi)P(Z \geq \phi) = \frac{\sqrt{N}}{2\pi} \int_{\phi}^{\infty} \int_{-\infty}^{\infty} W^2 \, e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2}$$

$$\cdot e^{-\frac{N}{2}\left[W - \frac{(N-1)\mu_1 + \mu_2 + \sqrt{\frac{N}{N-1}}\, Z}{N}\right]^2} dWdZ.$$

Integrating out the W, we obtain

$$E(W^2 | Z \geq \phi)P(Z \geq \phi) = \frac{N + [(N-1)\mu_1 + \mu_2]^2}{N^2} P(Z \geq \phi)$$

$$+ \frac{1}{N(N-1)\sqrt{2\pi}} \int_{\phi}^{\infty} Z^2 \, e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ$$

$$+ \frac{2[(N-1)\mu_1 + \mu_2]\sqrt{N}}{N^2\sqrt{2(N-1)\pi}} \int_{\phi}^{\infty} Z \, e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ. \qquad (36)$$

Combining (34), (35) and (36) we obtain

$$E(\bar{x}^{*2}) = \frac{N + [(N-1)\mu_1 + \mu_2]^2}{N^2}$$

$$+\frac{2[(N-1)\mu_1 + \mu_2]\sqrt{N}}{N^2\sqrt{2(N-1)\pi}} \int_{\phi}^{\infty} Z \; e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ$$

$$+ \frac{1}{N(N-1)\sqrt{2\pi}} \int_{\phi}^{\infty} Z^2 \; e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ. \tag{37}$$

Since

$$MSE(\bar{x}^*) = E(\bar{x}^{*2}) - E^2(\bar{x}^*) + (Bias)^2$$

and

$$E(\bar{x}^*) = \mu_1 + Bias,$$

it follows that

$$MSE(\bar{x}^*) = E(\bar{x}^{*2}) - 2\mu_1 E(\bar{x}^*) + \mu_1^2.$$

Substituting the values we obtained for $E(\bar{x}^{*2})$ and $E(\bar{x}^*)$, we have

$$MSE(\bar{x}^*) = \frac{1}{N} + \frac{(\mu_1 - \mu_2)^2}{N^2}$$

$$- \frac{2(\mu_1 - \mu_2)}{N\sqrt{2N(N-1)\pi}} \int_{\phi}^{\infty} Z \; e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]^2} dZ$$

$$+ \frac{1}{N(N-1)\sqrt{2\pi}} \int_{\phi}^{\infty} Z^2 \; e^{-\frac{1}{2}\left[Z - \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}}\right]} dZ. \tag{38}$$

When $\phi \longrightarrow -\infty$, $MSE(\bar{x}^*) \longrightarrow 1/(N-1)$; when $\phi \longrightarrow +\infty$, $MSE(\bar{x}^*) \longrightarrow 1/N + (\mu_1 - \mu_2)^2/N^2$. This serves as a partial check for (38).

To simplify the expression for the mean square error given by (38), let

$$Z = U + \frac{\sqrt{N-1}(\mu_1 - \mu_2)}{\sqrt{N}},$$

and integrate over the region

$$U + \frac{\sqrt{N-1}}{\sqrt{N}} \geq \phi,$$

where $\delta = \mu_1 - \mu_2 \geq 0$. We then have

$$MSE(\bar{x}^*) = \frac{N + \delta^2}{N^2} + \frac{\sqrt{N}\phi - \sqrt{N-1}\delta}{N(N-1)\sqrt{2N\pi}} e^{-\frac{1}{2}\left(\phi - \frac{\sqrt{N-1}\delta}{\sqrt{N}}\right)^2}$$

$$+ \frac{N - (N-1)\delta^2}{N^2(N-1)\sqrt{2\pi}} \int_{\phi - \frac{\sqrt{N-1}\delta}{\sqrt{N}}}^{\infty} e^{-\frac{U^2}{2}} dU. \qquad (39)$$

4. Derivation of $E(\bar{x}^*)$ and the mean square error of $\bar{x}^*$ (σ^2 unknown)

First we wish to find $E[(N-1)\bar{x}_{N-1} + x_N]/N$ if $t_1 < \phi$, where ϕ is the critical value of the t distribution corresponding to some preassigned significance level, say α_1. Let

$$y_1 = \frac{(\bar{x}_{N-1} - \mu_1)\sqrt{N-1}}{\sigma},$$

$$y_2 = \frac{(x_N - \mu_2)}{\sigma},$$

and

$$W = \frac{(N-2)s^2}{\sigma^2}.$$

Then the joint distribution of y_1, y_2, W is given by

$$\frac{1}{2^{\frac{N}{2}}\pi\,\Gamma\left(\frac{N-2}{2}\right)} e^{-\frac{y_1^2 + y_2^2}{2}} e^{-\frac{W}{2}} W^{\frac{N-4}{2}}. \tag{40}$$

Consider the following transformation:

$$u = \sqrt{\frac{N-1}{N}}\left(-\frac{y_1}{\sqrt{N-1}} + y_2\right),$$

$$v = \sqrt{\frac{N-1}{N}}\left(y_1 + \frac{y_2}{\sqrt{N-1}}\right),$$

$$W = W.$$

Then

$$y_1 = \sqrt{\frac{N-1}{N}}\left(v - \frac{u}{\sqrt{N-1}}\right),$$

$$y_2 = \sqrt{\frac{N-1}{N}}\left(u + \frac{v}{\sqrt{N-1}}\right), \tag{41}$$

and the Jacobian of the transformation is 1. From (40) and (41) we obtain the joint distribution of u,v,W:

$$\frac{1}{2^{\frac{N}{2}}\pi\,\Gamma\left(\frac{N-2}{2}\right)} e^{-\frac{u^2 + v^2}{2}} W^{\frac{N-4}{2}} e^{-\frac{W}{2}},$$

where $-\infty < u < \infty$, $-\infty < v < \infty$, $W > 0$.

However,

$$\frac{(N-1)\bar{x}_{N-1} + x_N}{N} = A + Bv$$

and

$$\frac{\bar{x}_{N-1} - x_N}{s\sqrt{1 + 1/(N-1)}} = \frac{C - Du}{F\sqrt{W}},$$

where

$$A = [(N-1)\mu_1 + \mu_2]/N,\ B = \sigma/\sqrt{N},\ C = (\mu_1 - \mu_2)/\sigma\ ,$$

$$D = \sqrt{N/(N-1)} \text{ and } F = \sqrt{N/(N-1)(N-2)}.$$

We then write

$$E[\frac{(N-1)\bar{x}_{N-1}+x_N}{N} | t_1 < \phi]\ P(t_1 < \phi) = E[A+Bv | \frac{C-Du}{F\sqrt{W}} < \phi]\ P(\frac{C-Du}{F\sqrt{W}} < \phi)$$

$$= \iiint (A+Bv)\ f(u,v,W)dudvdW,$$

where the region of integration is given by

$$\frac{C-Du}{F\sqrt{W}} < \phi, \ -\infty < v < \infty.$$

Integrating out the v, we have

$$E[\frac{(N-1)\bar{x}_{N-1}+x_N}{N} | t_1 < \phi]P(t_1 < \phi) = \frac{(N-1)\mu_1 + \mu_2}{N} \iint_R W^{\frac{N-4}{2}}$$

$$\cdot\ e^{-\frac{u^2+W}{2}}\ dudW, \qquad (42)$$

where the region of integration, R, is given by

$$\frac{C-Du}{F\sqrt{W}} < \phi.$$

We must now find $E(\bar{x}_{N-1} | t_1 \geq \phi)\ P(t_1 \geq \phi)$. To evaluate

this we note that

$$\bar{x}_{N-1} = \frac{\sigma\mu_1 y_1}{\sqrt{N-1}} = \mu_1 + Bv - Hu,$$

where

$$H = \frac{\sigma}{\sqrt{N(N-1)}} \text{ and } B = \frac{\sigma}{\sqrt{N}}.$$

Thus

$$E(\bar{x}_{N-1} | t_1 \geq \phi)P(t_1 \geq \phi) = E(\mu_1 + Bv - Hu \left| \frac{C-Du}{F\sqrt{W}} \geq \phi \right.) \; P(\frac{C-Du}{F\sqrt{W}} \geq \phi)$$

$$= \iiint (\mu_1 + Bv - Hu) \, f(u,v,W) \, dudvdW.$$

Integrating out the v, we have

$$E(\bar{x}_{N-1} | t_1 \geq \phi)P(t_1 \geq \phi) = \frac{1}{2^{\frac{N-2}{2}} \sqrt{2\pi} \, \Gamma\left(\frac{N-2}{2}\right)} \iint_{R'} (\mu_1 - Hu) \, W^{\frac{N-4}{2}}$$

$$\cdot \, e^{-\frac{u^2+W}{2}} \, dudW, \quad (43)$$

where the region of integration, R', is given by

$$\frac{C-Du}{F\sqrt{W}} \geq \phi.$$

Combining (42) and (43), we have

$$E(\bar{x}^*) = \mu_1 + \frac{\mu_2 - \mu_1}{N} P(R) - \frac{\sigma}{\sqrt{N(N-1)}} I_{R'}(u), \quad (44)$$

where

$$P(R) = \frac{1}{2^{\frac{N-2}{2}}\sqrt{2\pi}\,\Gamma\left(\frac{N-2}{2}\right)} \int\int_R e^{-\frac{u^2+W}{2}} W^{\frac{N-4}{2}}\, dudW$$

and

$$I_{R'}(u) = \frac{1}{2^{\frac{N-2}{2}}\sqrt{2\pi}\,\Gamma\left(\frac{N-2}{2}\right)} \int\int_{R'} u\, e^{-\frac{u^2+W}{2}} W^{\frac{N-4}{2}}\, dudW.$$

When $\phi \longrightarrow -\infty$, $E(\bar{x}^*) \longrightarrow \mu_1$; when $\phi \longrightarrow \infty$, $E(\bar{x}^*) \longrightarrow [(N-1)\mu_1 + \mu_2]/N$. This serves as a partial check for (44).

Result (44) is consistent with a theorem given by Kitagawa (1950). He assumed a random sample of size n_1 from a normal population $N(\xi_1,\sigma^2)$, and a second random sample of size n_2 from a normal population $N(\xi_2,\sigma^2)$. Using $s_i^2 = \sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2/n_i$ as his estimate of σ^2, he found the $E(\bar{x}^*)$ and the variance of $\bar{x}^*$ by first finding the distribution of $\bar{x}^*$.

Let us now put (44) in a different form by writing $P(R)$ and $I_{R'}(u)$ in terms of summations. When ϕ is a positive number, $(C-Du)/F\sqrt{W} < \phi$ may be replaced by the limits $-\infty < u < \infty$, $W > (Du-C)^2/\phi^2F^2 = (u - \sqrt{2\lambda})^2(N-2)/\phi^2 = Q$, say, where $\lambda = (\mu_1 - \mu_2)^2(N-1)/2N\sigma^2$. Then

$$P(R) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} \left[\frac{1}{2^{\frac{N-2}{2}}\,\Gamma\left(\frac{N-2}{2}\right)} \int_Q^{\infty} W^{\frac{N-4}{2}}\, e^{-\frac{W}{2}}\, dW \right] du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} \left[e^{-\frac{Q}{2}} \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{Q}{2}\right)^i \frac{1}{i!} \right] du,$$

where N is an even integer $\geq$ 4, and

$$I_{R'}(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u\, e^{-\frac{u^2}{2}} \left[\frac{1}{2^{\frac{N-2}{2}} \Gamma\left(\frac{N-2}{2}\right)} \int_0^Q W^{\frac{N-4}{2}} e^{-\frac{W}{2}} dW \right] du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u\, e^{-\frac{u^2}{2}} \left[1 - e^{-\frac{Q}{2}} \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{Q}{2}\right)^i \frac{1}{i!} \right] du,$$

for N an even integer $\geq$ 4.

Substituting these values into (44), we have

$$E(\bar{x}^*) = \mu_1 + \frac{\mu_2 - \mu_1}{N} \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} \left[e^{-\frac{Q}{2}} \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{Q}{2}\right)^i \frac{1}{i!} \right] du \right\}$$

$$+ \frac{\sigma}{\sqrt{N(N-1)}} \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u\, e^{-\frac{u^2}{2}} \left[e^{-\frac{Q}{2}} \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{Q}{2}\right)^i \frac{1}{i!} \right] du \right\}.$$

Using $Q = (u - \sqrt{2\lambda})^2 (N-2)/\phi^2$, we may write $E(\bar{x}^*)$ in the following form.

$$E(\bar{x}^*) = \mu_1 + \frac{\mu_2 - \mu_1}{N} \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{N-2}{2\phi^2}\right)^i \frac{1}{i!} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2+Q}{2}} (u$$

$$- \sqrt{2\lambda})^{2i}\, du \Big] + \frac{\sigma}{\sqrt{N(N-1)}} \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{N-2}{2\phi^2}\right)^i \frac{1}{i!} \Big[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u\, e^{-\frac{u^2+Q}{2}}$$

$$\cdot (u - \sqrt{2\lambda})^{2i}\, du \Big].$$

To simplify the integrals in the above expression, let $Z = u - \sqrt{2\lambda}$. Then

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2+Q}{2}} (u - \sqrt{2\lambda})^{2i} du = \frac{\phi}{\sqrt{\phi^2 + N - 2}}\, e^{-\frac{(N-2)\lambda}{\phi^2 + N - 2}}$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u\, e^{-\frac{u^2+Q}{2}} (u - \sqrt{2\lambda})^{2i} du = \frac{\phi\, \mu'_{2i+1}}{\sqrt{\phi^2 + N - 2}}\, e^{-\frac{(N-2)\lambda}{\phi^2 + N - 2}}$$

$$+ \frac{\phi\, \mu'_{2i}\, \sqrt{2\lambda}}{\sqrt{\phi^2 + N - 2}}\, e^{-\frac{(N-2)\lambda}{\phi^2 + N - 2}}$$

where μ'_{2i+1} and μ'_{2i} are the (2i+1)th and (2i)th moments about the origin of a normal distribution with mean

$$-\frac{\sqrt{2\lambda}\ \phi^2}{\phi^2+N-2}$$

and variance

$$\frac{\phi^2}{\phi^2+N-2}.$$

We now have

$$E(\bar{x}^*) = \mu_1 + \frac{\sigma\, e^{-\frac{(N-2)\lambda}{\phi^2 + N - 2}}}{\sqrt{N(N-1)(\phi^2 + N - 2)}} \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{N-2}{2\phi^2}\right)^i \frac{1}{i!}\, \mu'_{2i+1}$$

for N an even integer ≥ 4 and ϕ a positive number. As a partial check we note that when $\phi \longrightarrow \infty$, $E(\bar{x}^*) \longrightarrow [(N-1)\mu_1 + \mu_2]/N$.

When ϕ is a negative number, the region of integration $(C-Du)/F\sqrt{W} < \phi$ may be replaced by the limits $-\infty < u < \infty$ and $W < Q$. We then have

$$E(\bar{x}^*) = \frac{(N-1)\mu_1 + \mu_2}{N} - \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{N-2}{2\phi^2}\right)^i \frac{1}{i!} \left[\frac{\sigma(\mu'_{2i+1} + \sqrt{2\lambda}\, \mu'_{2i})}{\sqrt{N(N-1)}}\right.$$

$$\left. + \frac{\mu_2 - \mu_1}{N}\right] \frac{\phi}{\sqrt{\phi^2 + N - 2}}\, e^{\frac{-(N-2)\lambda}{\phi^2 + N - 2}} \qquad (45)$$

for N an even integer ≥ 4 and ϕ a negative number. As a partial check we note that when $\phi \longrightarrow -\infty$, $E(\bar{x}^*) \longrightarrow \mu_1$.

We now wish to find the mean square error of $\bar{x}^*$. Since

$$MSE(\bar{x}^*) = E(\bar{x}^{*2}) - E^2(\bar{x}^*) + (Bias)^2,$$

we need only find $E(\bar{x}^{*2})$ and subtract from it the terms of $E^2(\bar{x}^*)$ which are not contained in $(Bias)^2$. To find $E(\bar{x}^{*2})$, we write

$$E(\bar{x}^{*2}) = E[(A+Bv)^2 \mid \frac{C-Du}{F\sqrt{W}} < \phi]\, P(\frac{C-Du}{F\sqrt{W}} < \phi) + E[(\mu_1+Bv-Hu)^2 \mid$$

$$\frac{C-Du}{F\sqrt{W}} \geq \phi \Big] \; P\left(\frac{C-Du}{F\sqrt{W}} \geq \phi\right)$$

$$= \iiint (A+Bv)^2 \; f(u,v,W)\,du\,dv\,dW$$

$$+ \iiint (\mu_1+Bv-Hu)^2 \; f(u,v,W)\; du\,dv\,dW,$$

where f(u,v,W) is given on page 64. Integrating out the v, we have

$$E(\bar{x}^{*2}) = k\left[\iint_{R'} (\mu_1^2 + B^2 + H^2u^2 - 2H\,\mu_1)\; e^{-\frac{u^2+W}{2}}\; W^{\frac{N-4}{2}}\; du\,dW\right.$$

$$\left. + \iint_{R} (A^2 + B^2)\; e^{-\frac{u^2+W}{2}}\; W^{\frac{N-4}{2}}\; du\,dW\right],$$

where

$$k = \frac{1}{2^{\frac{N-2}{2}}\sqrt{2\pi}\;\Gamma\left(\frac{N-2}{2}\right)}$$

and the regions of integration are given by

$$R\left[\begin{array}{l} -\infty < u < \infty \\ \qquad W > Q \end{array}\right. \qquad R'\left[\begin{array}{l} -\infty < u < \infty \\ \qquad W < Q \end{array}\right.$$

when ϕ is positive.

Therefore,

$$E(\bar{x}^{*2}) = \mu_1^2 + B^2 + (A^2 - \mu_1^2)\; P(R) + H^2\; I_{R'}(u^2)$$

$$- 2H\;\mu_1\; I_{R'}(u),$$

where

$$I_{R'}(u^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 \, e^{-\frac{u^2}{2}} \left[\frac{1}{2^{\frac{N-2}{2}} \Gamma\left(\frac{N-2}{2}\right)} \int_0^Q e^{-\frac{W}{2}} W^{\frac{N-4}{2}} dW \right] du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 \, e^{-\frac{u^2}{2}} \left[1 - e^{-\frac{Q}{2}} \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{Q}{2}\right)^i \frac{1}{i!} \right] du$$

$$= 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 \, e^{-\frac{u^2+Q}{2}} \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{Q}{2}\right)^i \frac{1}{i!} \, du,$$

or

$$E(\bar{x}^{*2}) = \mu_1^2 + B^2 + H^2 + (A^2 - \mu_1^2 - H^2 \mu'_{2i+2} + 2H\mu'_{2i}$$

$$- 2H^2\sqrt{2\lambda}\, \mu'_{2i+1} - 2H^2 \lambda\, \mu'_{2i} + 2H\mu_1 \mu'_{2i+1})$$

$$\cdot \sum_{i=0}^{\frac{N-4}{2}} \left(\frac{N-2}{2\phi^2}\right)^i \frac{1}{i!} \frac{\phi}{\sqrt{\phi^2 + N - 2}} \, e^{-\frac{(N-2)\lambda}{\phi^2 + N - 2}}$$

for N an even integer ≥ 4 and ϕ positive.

We then have the following expression for $MSE(\bar{x}^*)$.

$$MSE(\bar{x}^*) = \frac{\sigma^2}{N-1} + \sum_{i=0}^{\frac{N-4}{2}} (A^2 - \mu_1^2 - H^2 \mu'_{2i+2} - 2H^2 \lambda\, \mu'_{2i}$$

$$- 2\, H^2\sqrt{2\lambda}\, \mu'_{2i+1} + 2H\mu_1\sqrt{2\lambda}\, \mu'_{2i}) \frac{1}{i!} \left(\frac{N-2}{2\phi^2}\right)^i$$

$$\cdot \frac{\phi}{\sqrt{\phi^2 + N - 2}} \, e^{-\frac{(N-2)\lambda}{\phi^2 + N - 2}} \tag{46}$$

for N an even integer ≥ 4 and ϕ positive.

If we take the limit as $\phi \longrightarrow \infty$, then $MSE(\bar{x}^*)$ approaches the value $[N\sigma^2 + (\mu_1 - \mu_2)^2]/N^2$. This serves as a partial check.

5. Test for H_0: $\mu_1 = \mu_0$ versus H_a: $\mu_1 \neq \mu_0$ (σ^2 known)

Let us now consider the hypothesis H_0 mentioned in Part 1 of Section C. We assume that σ^2 is known and, since there will be no loss of generality, we again take $\sigma^2 = 1$. The following test procedure is used:

(i) If $Z_1 \geq \phi$, test the hypothesis H_0: $\mu_1 = \mu_0$ versus H_a: $\mu_1 \neq \mu_0$ by applying the normal test to the statistic

$$Z_2 = \frac{\bar{x}_{N-1} - \mu_0}{\sqrt{1/(N-1)}}$$

with significance level α_2.

(ii) If $Z_1 < \phi$ and assuming $\mu_1 = \mu_2 = \mu_{12}$, say, test the hypothesis H_0: $\mu_{12} = \mu_0$ versus H_a: $\mu_{12} \neq \mu_0$ by applying the normal test to

$$Z_3 = \frac{\bar{x}_{12} - \mu_0}{\sqrt{1/N}}$$

with significance level α_3, where $\bar{x}_{12} = [(N-1)\bar{x}_{N-1} + x_N]/N$.

6. Power of the test procedure given in Part 5

The power of a test is the probability that a particular false hypothesis will be rejected. It is also defined as one minus the probability of committing a type II error where a type II error is the error committed by accepting a false hypothesis. We obtain the power P as the sum of two mutually exclusive components corresponding to the mutually exclusive alternatives given in (i) and (ii) of Part 5. We then have

$$P = \Pr(Z_1 \geq \phi \text{ and } |Z_2| \geq \phi_2) + \Pr(Z_1 < \phi \text{ and } |Z_3| \geq \phi_3)$$

or

$$P = P_1 + P_2,$$

where $P_1 = \Pr(Z_1 \geq \phi \text{ and } |Z_2| \geq \phi_2)$, $P_2 = \Pr(Z_1 < \phi \text{ and } |Z_3| \geq \phi_3)$ and ϕ, ϕ_2, and ϕ_3 are the critical values of the normal distribution corresponding to the significance levels α_1, α_2 and α_3, respectively.

The first term of the power, P_1, is obtained from the joint distribution of

$$Z = \frac{\bar{x}_{N-1} - x_N}{\sqrt{N/(N-1)}},$$

$$V = \frac{\bar{x}_{N-1} - \mu_0}{\sqrt{1/(N-1)}}.$$

Since Z and V are normally distributed variables with unit variances and means $\delta_1 = (\mu_1 - \mu_2)/\sqrt{N/(N-1)}$ and $\delta_2 = (\mu_1 - \mu_0)/\sqrt{1/(N-1)}$, respectively, and correlation coefficient $1/\sqrt{N}$, the joint distribution of Z and V is given by

$$f(Z,V) = \frac{\sqrt{N}}{2\pi\sqrt{N-1}} \; e^{-\frac{N[(Z-\delta_1)^2 + (V-\delta_2)^2 - \frac{2}{\sqrt{N}}(Z-\delta_1)(V-\delta_2)]}{2(N-1)}}.$$

Therefore, P_1 is the integral of the bivariate normal surface $f(Z,V)$ over the region $Z \geq \phi$, $|V| \geq \phi_2$. To table P_1, let $X = Z - \delta_1$, $Y = V - \delta_2$, $r = 1/\sqrt{N}$, and then use the tables given by K. Pearson (1931). In Table 5 below, the values of P_1 are given for various values of δ_1 and δ_2 when both ϕ and ϕ_2 are equal to 1.90 and $r = 0.5$.

Table 5. Values of P_1 for $\phi = \phi_2 = 1.90$, $r = 0.5$

	δ_1			
δ_2	0	1	3	5
0	0.0056	0.0177	0.0424	0.0571
1	0.0010	0.0774	0.1808	0.1859
3	0.0286	0.1803	0.7786	0.8641
5	0.0287	0.1841	0.8641	0.9981

The second term of the power, P_2, is obtained from the joint distribution of

$$Z = \frac{\bar{x}_{N-1} - x_N}{\sqrt{N/(N-1)}}$$

and

$$W = \frac{\bar{x}_{12} - \mu_0}{\sqrt{1/N}}.$$

Since Z and W are normally and independently distributed variables with unit variances and means δ_1 and $\delta_3 = [(N-1)\mu_1 + \mu_2 - N\mu_0]/\sqrt{N}$, their joint distribution is given by

$$f(Z,W) = \frac{1}{2\pi} e^{-\frac{1}{2}[(Z - \delta_1)^2 + (W - \delta_3)^2]}.$$

P_2 is the integral of the bivariate normal surface given by $f(Z,W)$ over the region $Z < \phi$, $|W| \geq \phi_3$ and may be evaluated as follows:

$$P_2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\phi} e^{-\frac{1}{2}(Z - \delta_1)^2} dZ \cdot \frac{1}{\sqrt{2\pi}} \int_{|W| \geq \phi_3} e^{-\frac{1}{2}(W - \delta_3)^2} dW$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\phi} e^{-\frac{1}{2}(Z - \delta_1)^2} dZ \left[1 - \frac{1}{\sqrt{2\pi}} \int_{-\phi_3}^{\phi_3} e^{-\frac{1}{2}(W - \delta_3)^2} dW \right].$$

Now, let $X = Z - \delta_1$, $Y = W - \delta_3$ and we have

$$P_2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\phi-\delta_1} e^{-\frac{X^2}{2}} dX \left[1 - \frac{1}{\sqrt{2\pi}} \int_{-\phi_3-\delta_3}^{\phi_3-\delta_3} e^{-\frac{Y^2}{2}} dY \right].$$

The value of P_2 is given in Table 6 for various values of δ_1 and δ_3 when both ϕ and ϕ_3 are equal to 1.90 and $r = 0.5$.

Table 6. Values of P_2 for $\phi = \phi_3 = 1.90$, $r = 0.5$

δ_3	δ_1 = 0	1	3	5
0	0.0558	0.0469	0.0078	0.0001
1	0.1806	0.1517	0.0252	0.0002
3	0.8395	0.7052	0.1174	0.0008
5	0.9703	0.8152	0.1355	0.0010

Since the power of the test is given by $P = P_1 + P_2$, we have

$$P = \frac{1}{2\pi\sqrt{1-r^2}} \int_{\phi-\delta_1}^{\infty} \int_{|U+\delta_2|\geq\phi_2} e^{-\frac{1}{2(1-r^2)}(X^2 + U^2 - 2rXU)} dUdX$$

$$+ \frac{1}{2\pi} \int_{-\infty}^{\phi-\delta_1} \int_{|Y+\delta_3|\geq\phi_3} e^{-\frac{1}{2}(X^2 + Y^2)} dYdX, \qquad (47)$$

where $X = Z - \delta_1$, $U = V - \delta_2$, $Y = W - \delta_3$ and $r = \sqrt{1/N}$.

Combining Tables 5 and 6, we have the following table of values for the power function, P, when $\phi = \phi_2 = \phi_3 = 1.90$ and $N = 4$ (i.e. $r = 0.5$).

Table 7. Values of P for $\phi = \phi_2 = \phi_3 = 1.90$, $r = 0.5$

	δ_2							
	0	1	3	5	0	1	3	5
δ_3								
	$\delta_1 = 0$				$\delta_1 = 1$			
0	0.0614	0.0568	0.0844	0.0845	0.0646	0.1243	0.2272	0.2310
1	0.1862	0.1816	0.2092	0.2093	0.1694	0.2291	0.3320	0.3358
3	0.8451	0.8405	0.8681	0.8682	0.7229	0.7826	0.8855	0.8893
5	0.9759	0.9713	0.9989	0.9990	0.8329	0.8926	0.9955	0.9993
	$\delta_1 = 3$				$\delta_1 = 5$			
0	0.0502	0.1886	0.7864	0.8719	0.0572	0.1860	0.8642	0.9982
1	0.0676	0.2060	0.8038	0.8893	0.0573	0.1861	0.8643	0.9983
3	0.1597	0.2981	0.8959	0.9814	0.0578	0.1867	0.8649	0.9989
5	0.1779	0.3163	0.9141	0.9996	0.0581	0.1869	0.8651	0.9991

In the special case when the null hypothesis H_0: $\mu_1 = \mu_0$, is true, the power is equal to the size of the test, i.e. to the probability of a type I error. When the null hypothesis is true, we have $\delta_1 = \delta_1$, $\delta_2 = 0$ and $\delta_3 = -\delta_1/\sqrt{N-1}$. Then

$$\text{Size of test} = \frac{1}{2\pi}\int_{-\infty}^{\phi-\delta_1}\int_{|Y-\frac{1}{\sqrt{N-1}}|>\phi_3} e^{-\frac{1}{2}(X^2+Y^2)}\,dY\,dX$$

$$+\frac{1}{2\pi\sqrt{1-r^2}}\int_{\phi-\delta_1}^{\infty}\int_{|U|>\phi_2} e^{-\frac{1}{2(1-r^2)}(X^2+U^2-2rXU)}\,dU\,dX.$$

The size of the test when $\phi = \phi_2 = \phi_3 = 1.90$ and $r = 0.5$ is given in Table 8 for various values of δ_1.

Table 8. Size of the test when $\phi = \phi_2 = \phi_3 = 1.90$, $r = 0.5$

N	δ_1 = 0	1	3	5
4	0.0614	0.0993	0.1171	0.0579

7. Test for $H_0: \mu_1 = \mu_0$ versus $H_a: \mu_1 \neq \mu_0$ after a preliminary test for an outlying observation has been made (σ^2 unknown)

When σ^2 is unknown, we use the following procedure:

(i) If $t_1 \geq t_{N-2,\alpha_1} = \phi$, test the hypothesis $H_0: \mu_1 = \mu_0$ versus $H_a: \mu_1 \neq \mu_0$ by applying the t test to the statistic

$$t_2 = \frac{\bar{x}_{N-1} - \mu_0}{s\sqrt{1/(N-1)}}$$

with N-2 degrees of freedom and significance level α_2, where $s^2 = \sum_{i=1}^{N-1} (x_i - \bar{x}_{N-1})^2/(N-2)$.

(ii) If $t_1 < t_{N-2,\alpha_1} = \phi$, then assuming $\mu_1 = \mu_2 = \mu_{12}$, say, test the hypothesis $H_0: \mu_{12} = \mu_0$ versus $H_a: \mu_{12} \neq \mu_0$ by applying the t test to the statistic

$$t_3 = \frac{\bar{x}_{12} - \mu_0}{s\sqrt{1/N}}$$

with N-2 degrees of freedom and significance level α_3, where $\bar{x}_{12} = [(N-1)\bar{x}_{N-1} + x_N]/N$.

8. Power of the test procedure given in Part 7

We obtain the power, P, as the sum of two mutually exclusive components corresponding to the mutually exclusive alternatives given in (i) and (ii) of Part 7. We have

$$P = Pr(t_1 \geq \phi \text{ and } |t_2| \geq \phi_2) + Pr(t_1 < \phi \text{ and } |t_3| \geq \phi_3)$$

or

$$P = P_1 + P_2,$$

where $P_1 = Pr(t_1 \geq \phi \text{ and } |t_2| \geq \phi_2)$, $P_2 = Pr(t_1 < \phi \text{ and } |t_3| \geq \phi_3)$ and ϕ, ϕ_2 and ϕ_3 are the critical values of the t distribution corresponding to significance levels α_1, α_2, and α_3, respectively. The first term of the power, P_1, is obtained from the joint distribution of $U = \bar{x}_{N-1} - x_N$, $V = \bar{x}_{N-1} - \mu_0$ and $W = (N-2)s^2/\sigma^2$. U and V are both normally distributed with means $\mu_1 - \mu_2$ and $\mu_1 - \mu_0$, variances $N\sigma^2/(N-1)$ and $\sigma^2/(N-1)$ and correlation coefficient $1/\sqrt{N}$. W is distributed as a χ^2 with N-2 degrees of freedom. Therefore, the joint distribution of U,V, W is given by

$$f(U,V,W) = \frac{\sqrt{N-1}}{2^{\frac{N-2}{2}} \pi \sigma^2 \Gamma\left(\frac{N-2}{2}\right)} \; e^{-\frac{W+S}{2}} \; W^{\frac{N-4}{2}},$$

where

$$S = \frac{1}{\sigma^2}\left[(U - \delta_1)^2 + N(V - \delta_2)^2 - 2(U - \delta_1)(V - \delta_2)\right],$$

$$\delta_1 = \mu_1 - \mu_2,$$

$$\delta_2 = \mu_1 - \mu_0.$$

Let us make the transformation of variables

$$T_1 = \frac{U}{\sqrt{\frac{WN\sigma^2}{(N-1)(N-2)}}},$$

$$T_2 = \frac{V}{\sqrt{\frac{W\sigma^2}{(N-1)(N-2)}}},$$

$$W = W,$$

then

$$U = \sqrt{\frac{WN\sigma^2}{(N-1)(N-2)}}\, T_1,$$

$$V = \sqrt{\frac{W\sigma^2}{(N-1)(N-2)}}\, T_2,$$

and the Jacobian of the transformation is

$$J = \frac{\sqrt{N}\, W\sigma^2}{(N-1)(N-2)}.$$

The distribution of the transformed variates is given by

$$f(T_1,T_2,W) = K\, W^{\frac{N-2}{2}}\, e^{-\frac{1}{2\sigma^2}(AW - 2BW^{\frac{1}{2}} + C)},$$

where

$$K = \frac{\sqrt{N}}{2^{\frac{N-2}{2}}\, \pi\, \sqrt{N-1}(N-2)\, \Gamma\left(\frac{N-2}{2}\right)},$$

$$A = \frac{N\sigma^2}{(N-1)(N-2)}\left[T_1^2 + T_2^2 - \frac{2}{\sqrt{N}}\, T_1T_2 + \frac{(N-1)(N-2)}{N}\right],$$

$$B = \frac{\sqrt{N}\,\sigma}{\sqrt{(N-1)(N-1)}}\left[(\delta_1 - \delta_2)T_1 + (\delta_2\sqrt{N} - \frac{\delta_1}{\sqrt{N}})T_2\right],$$

$$C = (\delta_1 - \delta_2)^2 + (N-1)\delta_2^2.$$

We then have

$$P_1 = \int_{\phi}^{\infty} \int_{|T_2| \geq \phi_2} \int_0^{\infty} f(T_1, T_2, W)\, dW dT_2 dT_1$$

or

$$P_1 = K \int_{\phi}^{\infty} \int_{|T_2| \geq \phi_2} \int_0^{\infty} e^{-\frac{1}{2\sigma^2}(C - \frac{B^2}{A})} e^{-\frac{A}{2\sigma^2}(W^{\frac{1}{2}} - \frac{B}{A})^2} \cdot W^{\frac{N-2}{2}}\, dW dT_2 dT_1.$$

However,

$$\int_0^{\infty} W^{\frac{N-2}{2}} e^{-\frac{A}{2\sigma^2}(W^{\frac{1}{2}} - \frac{B}{A})^2} dW = K' \sum_{i=0}^{N-1} \binom{N-1}{i} (-D)^{N-i-1} \cdot \int_D^{\infty} y^i e^{-y^2}\, dy,$$

where

$$K' = \frac{2^{\frac{N+2}{2}}\,\sigma^N}{A^{\frac{N}{2}}},\quad D = \frac{-B}{\sqrt{2A\,\sigma^2}},\quad y = \sqrt{\frac{A}{2\sigma^2}}(W^{\frac{1}{2}} - \frac{B}{A}),\ \text{and}$$

$$\int_D^\infty y^i e^{-y^2} dy = (1-s)[1 - \Phi(D)] \frac{(1;2;r)\sqrt{\pi}}{2^{r+1}}$$

$$+ e^{-D^2} \sum_{\nu=0}^{r-1} \frac{(i-1;-2;\nu)}{2^{\nu+1}} D^{i-1-2\nu},$$

where

$$i = 2r - s,\ s = 0 \text{ or } 1,\ (m;-d;\nu) = d^\nu\, \Gamma(\tfrac{m}{d}+1)/\Gamma(\tfrac{m}{d}+1-\nu),$$

$$(m;d;0) = 1,\ (m;d;\nu) = m(m+d)(m+2d) \ldots (m+2[\nu-1]),$$

$$\Phi(D) = \frac{2}{\sqrt{\pi}} \int_0^D e^{-t^2} dt.$$

Therefore, we have

$$P_1 = K\, 2^{\frac{N+2}{2}}\, \sigma^N \int_\phi^\infty \int_{|T_2| \geq \phi_2} e^{-\frac{1}{2\sigma^2}(C - \frac{B^2}{A})} A^{-\frac{N}{2}} \sum_{i=0}^{N-1} \binom{N-1}{i}$$

$$\cdot \left\{ (1-s)[1 - \Phi(D)] \frac{(1;2;r)\sqrt{\pi}}{2^{r+1}} + e^{-D^2} \sum_{\nu=0}^{r-1} \frac{(i-1;-2;\nu)}{2^{\nu+1} D^{1-i+2\nu}} \right\}$$

$$\cdot (-D)^{N-i-1}\ dT_2\, dT_1.$$

To obtain the second part of the power, P_2, we need the joint distribution of $U = \bar{x}_{N-1} - x_N$, $V' = \bar{x}_{12} - \mu_0$ and $W = (N-2)s^2/\sigma^2$. Since U and V' are independently and normally distributed variables with means $\delta_1 = \mu_1 - \mu_2$ and $\delta_3 = [(N-1)\mu_1 + \mu_2 - N\mu_0]/N$ and variances $N\sigma^2/(N-1)$ and σ^2/N, and W is distributed as a χ^2 with N-2 degrees of freedom, their joint distribution is given by

$$f(U,V',W) = \frac{\sqrt{N-1}}{2^{\frac{N}{2}}\pi\sigma^2\,\Gamma\left(\frac{N-2}{2}\right)} e^{-\frac{W}{2}} e^{-\frac{1}{2\sigma^2}\left[\frac{(N-1)(U-\delta_1)^2}{N}\right]} \cdot W^{\frac{N-4}{2}} e^{-\frac{1}{2\sigma^2}N(V'-\delta_3)^2}.$$

Now let

$$T_1 = \frac{U}{\sqrt{\frac{W\,N\,\sigma^2}{(N-1)(N-2)}}},$$

$$T_3 = \frac{V'}{\sqrt{\frac{W\,\sigma^2}{N(N-2)}}},$$

$$W = W,$$

then

$$U = \sqrt{\frac{W\,N\,\sigma^2}{(N-1)(N-2)}}\,T_1,$$

$$V' = \sqrt{\frac{W\,\sigma^2}{N(N-2)}}\,T_3,$$

$$W = W,$$

and the Jacobian of the transformation is

$$J = \frac{\sigma^2\,W}{\sqrt{N-1}\,(N-2)}.$$

The distribution of the transformed variates is given by

$$f(T_1,T_3,W) = K' \; W^{\frac{N-2}{2}} \; e^{-\frac{1}{2\sigma^2}(A'W - 2B'W^{\frac{1}{2}} + C')},$$

where

$$K' = \frac{1}{2^{\frac{N}{2}}\pi(N-2)\,\Gamma\left(\frac{N-2}{2}\right)},$$

$$A' = \sigma^2\left(\frac{T_1^2 + T_3^2}{N-2} + 1\right),$$

$$B' = \frac{\sigma\delta_1 T_1\sqrt{N-1}}{\sqrt{N(N-2)}} + \frac{\sigma\delta_3 T_3\sqrt{N}}{\sqrt{N-2}},$$

$$C' = \frac{(N-1)\,\delta_1^2}{N} + N\,\delta_3^2.$$

Then

$$P_2 = \int_{-\infty}^{\phi} \int_{|T_3| \geq \phi_3} \int_0^{\infty} f(T_1,T_3,W)\, dW dT_3 dT_1.$$

Integrating out the W, we have

$$P_2 = \sigma^N 2^{\frac{N}{2}} K' \int_{-\infty}^{\phi} \int_{|T_3| \geq \phi_3} e^{-\frac{1}{2\sigma^2}(C' - \frac{B'^2}{A'})} (A')^{-\frac{N}{2}}$$

$$\cdot \sum_{i=0}^{N-1} \binom{N-1}{i} (-D')^{N-i-1} \left\{ (1-s')[1 - \phi(D')] \frac{(1;2;r')\sqrt{\pi}}{2^{r'+1}} \right.$$

$$\left. + e^{-D'^2} \sum_{\nu'=0}^{r'-1} \frac{(i-1;-2;\nu')}{2^{\nu'+1}} (D')^{i-1-2\nu} \right\},$$

where

$$i = 2r' - s',\ s' = 0 \text{ or } 1,\ D' = \frac{-B'}{\sqrt{2A'\sigma^2}} \text{ and}$$

$$\Phi(D') = \frac{2}{\sqrt{\pi}} \int_0^{D'} e^{-t^2}\, dt.$$

The power is then given by $P_1 + P_2$. Integration difficulties prevent an explicit evaluation of the power of the test in this second case where σ^2 is unknown.

IV. LINE OUTLIER THEORY WITH SLOPE AND INTERCEPT CONSIDERED SEPARATELY

A. Introduction

We now wish to extend the univariate outlier methodology to the bivariate case. That is, we wish to construct an outlier methodology for straight lines. We term a line, $y=a+bx$, an outlier line if one of the following situations occurs:

(i) the slope b is an outlier, whatever the intercept,

(ii) the intercept a is an outlier, whatever the slope,

(iii) the line, considered in its entirety, is an outlier.

In the present chapter we deal primarily with case (i), the problem of slope outliers, and only briefly consider case (ii), the problem of intercept outliers. Discussion of case (iii), the entire line, is reserved for Chapter V.

Suppose that we have N sets of observations, the ith set consisting of n pairs (x_{ij},y_{ij}), $j = 1,2,\ldots,n$; $i = 1,2,\ldots,N$. The x_{ij} are known variables and the y_{ij} are random variables normally and independently distributed about their means, $\mu_i = \alpha_i + \beta_i x_{ij}$, with a common variance σ^2. We assume that the same set of x's is used in each of the N experiments. This is not too restrictive an assumption since experiments are often designed in this manner. We also assume that the true regression of y on x is linear for each set. Thus we are considering the model

$$y_{ij} = \alpha_i + \beta_i x_{ij} + e_{ij},$$

in which the e_{ij} are independently and normally distributed with mean zero and common variance. If the population lines have the same slope, then, of course, the β_i will all be equal.

The fitting of a straight line to a set of data is accomplished here by the method of least squares. The estimates of α_i and β_i are denoted by a_i and b_i, respectively, and the least squares criterion requires that they be chosen so as to minimize $\sum_j (y_{ij} - a_i - b_i x_{ij})^2$. The estimates of β_i and α_i are then $b_i = \sum_j (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)/\sum_j (x_{ij} - \bar{x}_i)^2$ and $a_i = \bar{y}_i - b_i \bar{x}_i$. It has been shown that confidence intervals may be constructed for the α_i and β_i by using the t distribution and that a joint confidence region for (α_i, β_i) may be obtained by using critical values of the F distribution.

The dispersion around the ith line, i.e. σ^2 is estimated by $s^2 = \sum_j (y_{ij} - Y_{ij})^2/(n-2)$, where Y_{ij} is the value obtained from the fitted line $Y_{ij} = a_i + b_i x_{ij}$ and y_{ij} is the corresponding observed value. We assume that the dispersion about each of the N lines is the same.

B. The Problem of Slope Outliers

Suppose that one of the b_i is considerably larger (or smaller) than the others. We would like to find a method for testing whether there is something unusual about this b_i, and hence whether it should be considered anomalous and rejected from the series. The present chapter consists of several somewhat different approaches to the problem, but in all cases

we wish to test the hypothesis that a particular slope (either the largest or the smallest) is not an outlier.

In order to handle the problem of outlying slopes in the most general way, the outlier test should be a test that takes into account the fact that a particular slope is to be tested, say the largest one. Several of the point outlier criteria discussed in Chapter II take this fact into account and modifications of these criteria are given in Sections C and D. The statistics V and V' proposed in Section E may be considered either hindsight statistics or statistics to be used in the special case where the scientist has inherited some additional information with the data.

The statistic V is one suggested by Rao (1952) for comparing the slopes of two lines. However, Rao did not explicitly derive the distribution of V, and since no derivation was found elsewhere, it is given explicitly in this thesis. The approach that leads to the V' statistic is based on overlapping confidence intervals and is somewhat analogous to one used by McCullough (1961) for another purpose. The distributional structure of the statistic is approximated by means of a device suggested by Patnaik (1949).

C. Modifications of Point Outlier Criteria

In Parts 1, 2 and 3 we apply three of the statistics mentioned in Chapter II to the following data in order to determine whether the slope, 0.8181, should be considered an

outlier. Use may be made of the respective modified point outlier criteria since the b_i are NID(β_i,$\sigma^2_{b_i}$), where $\sigma^2_{b_i} = \sigma^2 / \sum_j (x_{ij} - \bar{x}_i)^2$.

Given data:

x	y_1	y_2	y_3
-5	3	4	2
-4	3	3	2
-3	4	5	4
-2	6	7	4
-1	7	7	6
0	8	8	6
1	9	9	5
2	11	11	8
3	13	13	9
4	15	16	9
5	16	18	10

Results:

b	Error sum of squares
1.3818	4.5091
1.4363	13.2363
0.8181	5.2727

1. Extreme deviation statistic

The methods discussed in Chapter II seem to indicate that one of the obvious statistics to use to test for a single outlying slope is $(b_N - \bar{b})/\sigma_b$ [or $(\bar{b} - b_1)/\sigma$], where b_N (or b_1) is the suspected outlying slope, $\bar{b}$ is the mean of all the b_i and $\sigma^2_b = \sigma^2 / \sum_j (x_{ij} - \bar{x}_i)^2$. We make the assumptions that the data are from normal populations with homogeneous variability and that the same set of x's is used for each line.

The proposed statistic is similar to the u statistic proposed by McKay (1935). The following approximation to the u

distribution was given by McKay:

$$P_n(u) = \frac{n}{\sqrt{2\pi}} \int_{u\sqrt{n/(n-1)}}^{\infty} e^{-t^2/2} \, dt.$$

To apply the test statistic to the data given above, we make the further assumption that the value of σ^2 is closely approximated by s^2, where s^2 equals the pooled error sum of squares divided by the pooled degrees of freedom. We then have $(\bar{b} - b_1)/\sigma_b = (1.2124 - 0.8181)/0.0880 = 4.48$ and $P_3(4.48) \approx 0$. We conclude from this that the slope 0.8181 is an outlier.

2. The statistic S_N^2/S^2

This statistic, proposed and tabled by Grubbs (1950), obviates the necessity of estimating the population variance when it is unknown. A discussion of the statistic is given in Chapter II. To apply the statistic to the data of Section C, we need to calculate

$$S_N^2/S^2 = \sum_{i=2}^{3} (b_i - \bar{b}_1)^2 / \sum_{i=1}^{3} (b_i - \bar{b})^2,$$

where the numerator excludes the suspected outlying slope, 0.8181, from the mean $\bar{b}_1$ and from the summation. We obtain $S_N^2/S^2 = 0.0063$ and conclude that the value 0.8181 is an outlier, since the significance point at the 5% level is 0.0027.

3. Dixon's statistic

To apply Dixon's statistic we calculate r_{11}, i.e. $(b_2 - b_1)/(b_3 - b_1) = 0.912$. Using Dixon's table we reject the hypothesis and conclude that $b = 0.8181$ is an outlier.

D. The Problem of Intercept Outliers

The model that we assume in this section is

$$y_{ij} = \alpha_i^* + \beta_i(x_{ij} - \bar{x}_i) + e_{ij},$$

where the x_{ij} are known variables and the y_{ij} are again random variables normally and independently distributed about their means with a common variance. We also assume that the same set of x's is used in each of the N experiments.

Let us assume that we have a set of N values of a^* ($a^* = \hat{a}^* = \bar{y}$) and wish to test the hypothesis that these N values are from the same normally distributed population versus the alternative hypothesis that one of the a^* is from a normal population whose mean differs from the mean of the population generating the other a^*'s. Since the N values of a^* are actually the respective means of the y's for each line, we can apply the tests described in Chapter II to this problem. As an example, consider the application of Grubbs' statistic S_N^2/S^2, to the three a^* values in the problem in Section C, namely, 8.6363, 9.1818 and 5.9091. To test the value 5.9091, we calculate $S_N^2/S^2 = 0.024$, refer to Grubbs' table and conclude that 5.9091 is an outlier.

For a second example, consider the extreme deviation statistic. The form of the extreme deviation statistic in this situation is

$$\frac{\bar{y}_N - \bar{\bar{y}}}{\sigma/\sqrt{n}}$$

where $\bar{\bar{y}}$ is the mean of all the $\bar{y}_i$ and $\bar{y}_N$ is the suspected outlier.

E. Outlier Tests Assuming Available A Priori Information Sufficient to Identify a Suspected Slope Outlier

1. The construction of the statistic V

The following approach to the problem of outlying slopes is, as we mentioned previously, based on a statistic proposed by Rao (1952) for comparing two lines, and is, essentially, a hindsight statistic.

Assume that we have N groups of observations (x_{ij},y_{ij}), $j = 1,2,\ldots,n$; $i = 1,2,\ldots,N$. A separate line, $Y_{ij} = a_i + b_i x_{ij}$, $i = 1,2,\ldots,N$, can be fitted to each group of observations, and the variation about each line, s_i^2, can be obtained. We assume that there exists a common residual variance.

Let $Y_{oj} = a_o + b_o x_{oj}$ be the line containing the largest (or smallest) value for the slope. Let $Y_{aj} = a_a + b_a x_{aj}$ be the line fitted to all the data except that used to determine b_o. We use the following notation for the two sets of data used to obtain these lines:

Sample size

n_o n_a

Mean values

$\bar{x}_o, \bar{y}_o$ $\bar{x}_a, \bar{y}_a$

Corrected sums of squares and products

$$S_o = \sum_{j=1}^{n_o} (x_{oj}-\bar{x}_o)^2 \qquad S_a = \sum_{j=1}^{n_a} (x_{aj}-\bar{x}_a)^2$$

$$Q_o = \sum_{j=1}^{n_o} (x_{oj}-\bar{x}_o)(y_{oj}-\bar{y}_o) \qquad Q_a = \sum_{j=1}^{n_a} (x_{aj}-\bar{x}_a)(y_{aj}-\bar{y}_a)$$

Residual sum of squares

$$\sum_{j=1}^{n_o} (y_{oj}-\bar{y}_a)^2 - b_o Q_o \qquad \sum_{j=1}^{n_a} (y_{aj}-\bar{y}_a)^2 - b_a Q_a.$$

Then

$$R_0 = \sum_{j=1}^{n_o} (y_{oj}-\bar{y}_o)^2 - b_o Q_o + \sum_{j=1}^{n_a} (y_{aj}-\bar{y}_a)^2 - b_a Q_a$$

has n_o+n_a-4 degrees of freedom.

We consider the value b_o to be an outlier if the hypothesis H_0: $\beta_o=\beta_a$ is rejected. To test this hypothesis we calculate $S_1 = S_o+S_a$, $Q_1 = Q_o+Q_a$ and obtain b_1 from the equation $Q_1 = b_1 S_1$. The residual sum of squares

$$R_2 = \sum_{j=1}^{n_o} (y_{oj}-\bar{y}_o)^2 + \sum_{j=1}^{n_a} (y_{aj}-\bar{y}_a)^2 - b_1 Q_1$$

has n_o+n_a-3 degrees of freedom. The statistic used to test the hypothesis is given by

$$V = \frac{(R_2 - R_0)/1}{R_0/(n_o+n_a-4)}.$$

In Part 3 we show that, under the null hypothesis, V is distributed as an F with 1 and n_o+n_a-4 degrees of freedom. First, however, we consider a numerical example.

2. Numerical example

Let us calculate the value of V for the data given in Section C. We have

R_0 = residual sum of squares for the outlier line + residual sum of squares for the average line

= 5.273 + 26.200

= 31.473,

R_2 = total sum of squares for the outlier line + total sum of squares for the average line - b_1Q_1

= 78.909 + 442.727 - 475.200

= 46.436.

Then

$$V = \frac{(R_2 - R_0)/1}{R_0/(n_o+n_a-4)} = 13.7.$$

The critical value of the F distribution with 1 and 18 degrees of freedom at the 5% level is 4.41; therefore, we conclude that the slope b = 0.8181 is an outlier.

For this example, the additional a priori information mentioned in Section B might be that previous investigations

would lead one to suspect any slope less than one as a possible outlier.

3. Derivation of the frequency function of V

Theorem 1. Let V be a random variable of the form given in Part 2, where R_2-R_0 and R_0 are independent random variables. Then the frequency function of V is a central F with 1 and n_o+n_a-4 degrees of freedom if the null hypothesis is true, and a non-central F if the null hypothesis is not true.

The proof of this theorem follows from two lemmas:

Lemma 1. $(R_2-R_0)/\sigma^2$ is distributed as a χ^2 with one degree of freedom if the null hypothesis is true and a non-central χ^2 with one degree of freedom if the alternative hypothesis is true.

proof: We may write

$$R_2-R_0 = \frac{S_o^2Q_a^2 + S_a^2Q_o^2 - 2S_oS_aQ_oQ_a}{S_oS_a(S_o+S_a)}$$

or

$$(R_2-R_0)/\sigma^2 = \frac{(S_oQ_a - S_aQ_o)^2}{\sigma^2S_oS_a(S_o+S_a)}$$

$$= T^2,$$

where $T = (S_oQ_a - S_aQ_o)/\sigma\sqrt{S_oS_a(S_o+S_a)}$ is a linear combination of normally distributed variables and, therefore, is itself normally distributed.

Now

$$E(T) = S_o S_a(\beta_a-\beta_o)/\sigma\sqrt{S_o S_a(S_o+S_a)},$$

$$E(T^2) = [S_o S_a(\beta_o-\beta_a)^2 + \sigma^2(S_o+S_a)]/\sigma^2(S_o+S_a)$$

and

$$Var(T) = 1.$$

Therefore, under the null hypothesis, T is distributed as a standard normal. Hence, T^2 is distributed as a χ^2 with 1 degree of freedom.

When H_0 is not true, $E(T)\neq 0$. In this case, T is distributed as $N[E(T),1]$, and T^2 is distributed as a non-central χ^2 with 1 degree of freedom and non-centrality parameter given by

$$\lambda = S_o S_a(\beta_o-\beta_a)^2/2\sigma^2(S_o+S_a),$$

and the frequency function is given by

$$f(T^2) = e^{-\lambda}\sum_{i=0}^{\infty} \frac{\lambda^i (T^2)^{(2i-1)/2} e^{-T^2/2}}{i!\ 2^{(2i+1)/2}\ \Gamma\left(\frac{1+2i}{2}\right)}; \quad T^2 > 0.$$

Lemma 2. R_0/σ^2 is always distributed as a χ^2 with n_o+n_a-4 degrees of freedom.

The proof of this lemma is given by Kenney and Keeping (1960).

4. Calculation of the expected mean squares

Since

$$R_2 = \sum_{j=1}^{n_o}(y_{oj}-\bar{y}_o)^2 + \sum_{j=1}^{n_a}(y_{aj}-\bar{y}_a)^2 - b_1 Q_1$$

and

$$R_0 = \sum_{j=1}^{n_o} (y_{oj}-\bar{y}_o)^2 + \sum_{j=1}^{n_a} (y_{aj}-\bar{y}_a)^2 - b_oQ_o - b_aQ_a,$$

we have

$$R_2-R_0 = b_oQ_o + b_aQ_a - b_1Q_1.$$

Therefore, in order to calculate $E(R_2-R_0)$, we must find $E(b_oQ_o)$, $E(b_aQ_a)$ and $E(b_1Q_1)$. First, consider $E(b_oQ_o)$.

$$\begin{aligned} E(b_oQ_o) &= E(Q_o^2/S_o) \\ &= E[\sum_{j=1}^{n_o} (x_{oj}-\bar{x}_o)(y_{oj}-\bar{y}_o)]^2/S_o \\ &= E[\beta_oS_o + \sum_{j=1}^{n_o} (x_{oj}-\bar{x}_o)(e_{oj}-\bar{e}_o)]^2/S_o \\ &= (\beta_o^2S_o^2 + \sigma^2S_o)/S_o \\ &= \beta_o^2S_o + \sigma^2. \end{aligned}$$

Similarly,

$$E(b_aQ_a) = \beta_a^2S_a + \sigma^2.$$

Now,

$$\begin{aligned} E(b_1Q_1) &= E[(Q_o+Q_a)^2/(S_o+S_a)] \\ &= [E(Q_o^2) + E(Q_a^2) + 2E(Q_oQ_a)]/(S_o+S_a) \\ &= [\beta_o^2S_o^2 + S_o\sigma^2 + \beta_a^2S_a^2 + \sigma^2S_a + 2\beta_oS_o\beta_aS_a]/(S_o+S_a). \end{aligned}$$

Therefore,

$$E(R_2-R_0) = \sigma^2 + S_oS_a(\beta_o-\beta_a)^2/(S_o+S_a).$$

It is obvious that the expected value of (R_2-R_0) is σ^2 whenever the null hypothesis is true.

In Part 3, we showed that R_0/σ^2 is always distributed as a χ^2 with n_o+n_a-4 degrees of freedom. Since the expectation of χ^2 is equal to the number of degrees of freedom, we have

$$E(R_0/\sigma^2) = n_o+n_a-4$$

and

$$E[R_0/(n_o+n_a-4)] = \sigma^2, \text{ always.}$$

5. Power of the test

We have seen that if $\beta_o-\beta_a = \delta \neq 0$, the numerator of V will be distributed as a non-central χ^2 with non-centrality parameter λ. Since the denominator of V is always distributed as a central χ^2, we know that when the null hypothesis is not true, V is distributed as a non-central F with 1 and n_o+n_a-4 degrees of freedom and non-centrality parameter λ. There are three parameters in the V distribution; the degrees of freedom for the chi-square in the numerator of V, the degrees of freedom for the chi-square in the denominator, and the non-centrality parameter of the chi-square in the numerator. Tables have been devised by Tang (1938) to evaluate

$$\int_0^{F_\alpha} f(F')\, dF'$$

for certain values of F_α, where F' is a non-central F. These tables are given in terms of E^2, where $E^2 = F'/(n_o+n_a-4+ F')$.

The frequency function for E is

$$g(E^2) = \sum_{i=0}^{\infty} \frac{\Gamma\left(\frac{2i+n_o+n_a-3}{2}\right) \lambda^i e^{-\lambda}}{\Gamma\left(\frac{n_o+n_a-4}{2}\right) \Gamma(i+\frac{1}{2})\, i!} (E^2)^{i-\frac{1}{2}} (1-E^2)^{(n_o+n_a-6)/2}$$

where $0 < E^2 < 1$.

The tables give the values of the probability of a type II error, i.e.

$$\int_0^{E_\alpha^2} g(E^2)\, dE^2,$$

where E^2 is determined from

$$\int_{E_\alpha^2}^{1} f(E^2 | \lambda=0)\, dE^2 = \alpha,$$

α being chosen as either 0.01 or 0.05.

In Tang's notation, we have

$$P(II) = 1 - \beta(\phi) = \int_0^{E_\alpha^2} g(E^2;\, f_1, f_2, \phi)\, dE^2,$$

where f_1 represents the numerator degrees of freedom, f_2 the degrees of freedom for the denominator and $\phi = \sqrt{2\lambda/(f_1+1)}$. In our problem, $f_1 = 1$, $f_2 = n_o+n_a-4$ and $\phi = \sqrt{\lambda}$.

Testing the hypothesis that $\lambda=0$ in the F' distribution is equivalent to testing our original hypothesis. To test the hypothesis that $\lambda=0$, we use the interval $F_\alpha < F' < \infty$ as the critical region of size α. The power of the test, $\beta(\lambda)$, is the probability that the observed F' falls in the critical region when $\lambda \neq 0$, and is given by

$$\beta(\lambda) = \int_{F_\alpha}^{\infty} f(F')\, dF'$$

or

$$\beta(\lambda) = \int_{E_\alpha^2}^{1} g(E^2)\, dE^2.$$

Therefore, the power of the test is given by

$$\beta(\lambda) = 1 - P(II).$$

6. Special case

In this section we again assume that the same set of x's is used in each of the N experiments and make the additional assumption that $\bar{x}_i = 0$. We then have $n_o = n$, $n_a = n(N-1)$, $S_a = (N-1)S_o$, and $\lambda = [(N-1)S_o/2N](\delta/\sigma)^2$.

Example 1. Suppose that the x's chosen for each line are -2, -1,0,1 and 2 and that we have six lines. Then $n = 5$, $N = 6$, $n_a = 25$, $S_a = 10$, $f_1 = 1$, $f_2 = 26$, $\lambda = (25/6)(\delta/\sigma)^2$, and $\phi = (5/6)(\delta/\sigma)$. The power for various values of δ/σ is given below for $\alpha = 0.01$.

δ/σ	0	1.0	1.2	1.4	1.6	1.8	2.0
	0	0.116	0.558	0.744	0.810	0.939	0.997

Example 2. Consider the same x_i as in Example 1, but let the number of lines be 13. Then $\lambda = (60/13)(\delta/\sigma)^2$, and $\phi = 2.148(\delta/\sigma)$. The power for various values of δ/σ is given below for $\alpha = 0.01$.

δ/σ	0	1.0	1.2	1.4	1.6	1.8	2.0
	0	0.641	0.827	0.940	0.966	0.990	0.999

7. The construction of the statistic V'

The following approach to the problem of outlying slopes is based on overlapping confidence intervals. First, we find the $100(1-\alpha)\%$ confidence interval for β_o, the regression coefficient suspected of being an outlier, and the $100(1-\alpha)\%$ confidence interval for β_a, the regression coefficient of the line described in Part 1. If the two confidence intervals fail to intersect, we reject the hypothesis that the regression coefficient is not an outlier.

We assume that the x_i are the same for each of the N lines, that $\sum_j x_{ij} = 0$ and that $\sum_j x_{ij}^2 = n$. Then the confidence interval for β_a,

$$b_a \pm t_{n_a-2,\alpha/2}\, s_{b_a},$$

may be written as

$$b_a \pm t_{n_a-2,\alpha/2}\, s_a/\sqrt{n_a} \tag{48}$$

and the confidence interval for β_o becomes

$$b_o \pm t_{n_o-2\ \alpha/2} s_o/\sqrt{n_o}. \tag{49}$$

Confidence intervals (48) and (49) will not intersect if $V' > 1$ or $V' < -1$, where V' is defined as

$$V' = \frac{b_a - b_o}{t_{n_a-2,\alpha/2}\, s_a/\sqrt{n_a} + t_{n_o-2,\alpha/2}\, s_o/\sqrt{n_o}}, \tag{50}$$

where $n_o = n$ and $n_a = n(N-1)$.

Therefore, the test which is made using V' is to reject

H_0: $\beta_o=\beta_a$ if $V'>1$ or $V'<-1$, and not to reject otherwise.

8. Numerical example

Let us calculate the value of V' for the data given in Section C. Transforming the x_i so that $\sum_j x_{ij}^2=n$, we find that $s_{b_o}^2=58/121$, $s_{b_a}^2=131/110$ and finally $V'=0.47$. Since V' is less than 1, we reject the hypothesis and conclude that b_o is an outlier.

9. The distributional structure of V'

Theorem 2. The frequency function of V' is of the form

$$\frac{N(0,1)}{(\phi_1 \chi_{n_a-2} + \phi_2 \chi_{n_o-2})},$$

where $\phi_1 = t_{n_a-2}/\sqrt{N(n_a-2)}$ and $\phi_2 = t_{n_o-2}\sqrt{N-1}/\sqrt{N(n_o-2)}$.

proof: Since both b_a and b_o are normally distributed with means β_a, β_o and variances σ^2/n_a, σ^2/n_o, respectively, we know that

$$\frac{(b_a-b_o) - (\beta_a-\beta_o)}{\sigma(1/n_a+1/n_o)^{1/2}} \tag{51}$$

is normally distributed with mean 0 and variance 1. Now consider

$$\frac{t_{n_a-2}s_a/\sqrt{n_a} + t_{n_o-2}s_o/\sqrt{n_o}}{\sigma(1/n_o+1/n_a)^{1/2}}. \tag{52}$$

Using the fact that $(n_a-2)s_a^2/\sigma^2$ is distributed as a chi-square with n_a-2 degrees of freedom, and $(n_o-2)s_o^2/\sigma^2$ is distributed

as a χ^2 with n_o-2 degrees of freedom, we may rewrite (52) as

$$\phi_1\chi_{n_a-2} + \phi_2\chi_{n_o-2},$$

where ϕ_1 and ϕ_2 are given above. Under the null hypothesis, V' may be written as the quotient of (51) and (52) and the result follows.

10. The approximate distribution of V'

We can approximate the quantity $(\phi_1\chi_{n_a-2} + \phi_2\chi_{n_o-2})$ by using $(\gamma\chi^2_\nu)^{1/2}$ where the constants γ and ν are found by equating the first two moments of $\gamma\chi^2_\nu$ and $(\phi_1\chi_{n_a-2} + \phi_2\chi_{n_o-2})^2$; the device suggested by Patnaik (1949).

The kth moment of χ^2_ν is $\nu(\nu+2)(\nu+4) \ldots (\nu+2[k-1])$. Thus the first moment of $\gamma\chi^2_\nu$ is $\gamma\nu$ and the second moment is $\gamma^2\nu(\nu+2)$. The kth moment of χ_n is

$$\frac{2^{k/2}\,\Gamma\left(\frac{n+k}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

Thus the first two moments of χ_n are $2^{1/2}\,\Gamma[(n+1)/2]/\,\Gamma(n/2)$ and n.

The first moment of $(\phi_1\chi_{n_a-2} + \phi_2\chi_{n_o-2})^2$ is

$$E(\phi_1\chi_{n_a-2} + \phi_2\chi_{n_o-2})^2 = \phi_1^2(n_a-2) + \phi_2^2(n_o-2) + 2b_1b_2\phi_1\phi_2,$$

where

$$b_1 = \frac{2^{1/2}\,\Gamma[(n_a-1)/2]}{\Gamma[(n_a-2)/2]} \quad \text{and} \quad b_2 = \frac{2^{1/2}\,\Gamma[(n_o-1)/2]}{\Gamma[(n_o-2)/2]}$$

The second moment of $(\phi_1 \chi_{n_a-2} + \phi_2 \chi_{n_o-2})^2$ is

$$E(\phi_1 \chi_{n_a-2} + \phi_2 \chi_{n_o-2})^4 = \phi_1^4(n_a-2)n_a + \phi_2^4(n_o-2)n_o + 6\phi_1^2\phi_2^2(n_a-2)(n_o-2) + 4\phi_1\phi_2 b_1 b_2[\phi_1^2(n_a-1) + \phi_2^2(n_o-1)].$$

Equating the first moments and then the second moments, we have

$$H_1 = \gamma \nu,$$
$$H_2 = \gamma^2 \nu(\nu+2),$$

or

$$\gamma = (H_2 - H_1^2)/2H_1,$$
$$\nu = 2H_1^2/(H_2 - H_1^2),$$

where H_1 and H_2 represent the first and second moments, respectively, of $(\phi_1 \chi_{n_a-2} + \phi_2 \chi_{n_o-2})^2$.

Therefore, the approximate distribution of V' is

$$\frac{N(0,1)}{(\gamma \chi_\nu^2)^{1/2}}$$

or

$$t_\nu / H_1^{1/2}.$$

11. Size of the test using the approximate distribution

The probability of a type I error is given by

$$P(t_\nu/\sqrt{H_1} > 1 \text{ or } < -1) = P(t_\nu > \sqrt{H_1} \text{ or } < -\sqrt{H_1}) = 1 - P(-\sqrt{H_1} < t_\nu < \sqrt{H_1}).$$

Thus, the size of the test depends upon (i) the value of α used in determining the confidence intervals for β_o and β_a, (ii) the number of lines in the experiment, and (iii) the number of points used to determine each line.

No attempt will be made in this thesis to discuss the power of the test.

V. LINE OUTLIER THEORY WITH SLOPE AND INTERCEPT CONSIDERED JOINTLY

A. Introduction

It is sometimes necessary to know whether several fitted regression lines are estimates of the same population line. The problem considered in this chapter concerns the possibility that one out of a set of N straight lines, considered in its entirety, differs radically from the remaining lines, i.e. one of them is an outlier line. We envision a research situation that requires that a line not be termed an outlier line unless both the slope and intercept considered jointly are outliers.

The general test criterion in point outlier theory is defined in terms of the distance between the largest observation and the mean of the observations. In order to formulate a line outlier methodology in an analogous manner, we need some measure of distance between an average line and a particular line, namely the line farthest from the average line in some distance sense.

In Section B each of the N lines is represented by a point whose coordinates correspond to the intercept and slope of the line. The distance between each of these points and an average point is determined and a statistic based on the ratio of the square of the largest of these N distances to the sum of the squares of all the distances is proposed. This

statistic takes into account the fact that we choose the point farthest from the average point to test as a possible outlier. We have mentioned several times that the practice of selecting a large value, not in advance, but because of its exceptional size requires a special test of significance. In other words, if we wish to test the significance of the largest observed distance we must compare the value observed with the sampling distribution of the largest of, say, N independent values, and not with that of any one value chosen in advance. Unfortunately, the exact distribution of the criterion proposed has not been determined. Hence, we propose using an approximation to this distribution. The approximation is based on a distribution given by Fisher (1929) and holds for large samples.

The statistic proposed in Section C is based on the fact that the slope and intercept form a bivariate normal distribution and hence the criterion proposed by Wilks (1963) for multivariate statistical outliers may be used. In Section D we discuss a criterion based on Siotani's (1959) statistic. An empirical comparison of the statistics given in Sections B, C and D is given in Section E, where we have generated ten lines from one population and one line from another population.

In Section F we propose test criteria which may be considered either as hindsight tests or as special cases where additional a priori information about the population is available along with the sample data. The statistic U given in Section F was suggested by Rao (1952) to test whether two re-

gression lines are the same; however, he did not explicitly derive its distribution and since no derivation was found elsewhere it is included in this thesis. The statistic U' given in Section F is based on overlapping confidence regions.

In brief then, this chapter consists of methods to be used when no additional information is available about the population other than the sample data, and methods to be employed when some a priori information is available.

B. A Test Criterion Based on Maximum Distance

Let us assume that we have N sets of observations (x_{ij}, y_{ij}), $i = 1,2,\ldots,N$; $j = 1,2,\ldots,n$, giving rise to N linear equations, where for each value of x_{ij}, y_{ij} is normally and independently distributed about $\alpha_i^* + \beta_i(x_{ij}-\bar{x}_i)$ with variance σ^2. If the population lines are identical, then $\alpha_i^* = \alpha^*$ and $\beta_i = \beta$ for all i. The N least squares estimates of α^* and β are denoted by a_i^* and b_i, respectively. We define

$$\bar{a} = \sum_{i=1}^{N} a_i^*/N,$$

$$\bar{b} = \sum_{i=1}^{N} b_i/N,$$

$$d_i^2 = (a_i^*-\bar{a})^2 + (b_i-\bar{b})^2, \quad i = 1,2,\ldots,N$$

$$r_i = d_i^2 / \sum_{i=1}^{N} d_i^2, \text{ and}$$

$$R = \max_i r_i.$$

$$= \frac{d^2_{max.}}{\sum_{i=1}^{N} d_i^2} = \frac{(a_r^*-\bar{a})^2 + (b_r-\bar{b})^2}{\sum_{i=1}^{N}[(a_i^*-\bar{a})^2 + (b_i-\bar{b})^2]},$$

where (a_r^*, b_r) represents the point farthest from $(\bar{a},\bar{b})$.

We propose the statistic R as a criterion for detecting one outlying line. The test procedure involves the following steps:

1. Calculate d_i^2, i=1,2,...,N.
2. Select the largest d_i^2, say $d^2_{max.} = (a_r^*-\bar{a})^2+(b_r-\bar{b})^2$.
3. Calculate
$$R = \frac{d^2_{max.}}{\sum_{i=1}^{N} d_i^2}.$$
4. Reject the line, $Y_{rj} = a_r^* + b_r(x_{rj}-\bar{x}_r)$ if R is too large, say $R > C$.

To determine the critical value C we must find the distribution of R. After these preparations, we now observe:

Theorem 3. If all the a_i^*'s estimate the population value α^*, all the b_i estimate the population value β, and $\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 = n$, then

(i) the quantities $v_1, v_2, \ldots, v_N$ [where $v_i = nNd_i^2/(N-1)$] are each distributed as $\chi^2\sigma^2$ with two degrees of freedom, and

(ii) the ratio of the largest of the v's to their total, i.e.

$$\frac{v_{max.}}{\sum_{i=1}^{N} v_i} = \frac{\max_i [(a_i^*-\bar{a})^2 + (b_i-\bar{b})^2]}{\sum_{i=1}^{N} [(a_i^*-\bar{a})^2 + (b_i-\bar{b})^2]}$$

$$= R,$$

and the probability that this ratio R exceeds the value R' is approximated by

$$N(1-R')^{N-1} - \frac{N(N-1)}{2!}(1-2R')^{N-1} + \ldots + (-1)^{k-1}\frac{N!}{k!(N-k)!}(1-kR')^{N-1}, \quad (53)$$

where k is the greatest integer less than $1/R'$.

proof: We have

$E(a_i^*) = E(\bar{y}_i) = \alpha^*$, $V(a_i^*) = V(\bar{y}_i) = \sigma^2/n$, $E(\bar{a}) = \alpha^*$, $V(\bar{a}) = \sigma^2/nN$, $E(b_i) = \beta$, $V(b_i) = \sigma^2/n$, $E(\bar{b}) = \beta$, $V(\bar{b}) = \sigma^2/nN$.

Then

$$\frac{(b_i-\bar{b})^2 nN}{(N-1)\ \sigma^2} \quad \text{and} \quad \frac{(a_i^*-\bar{a})^2 nN}{(N-1)\ \sigma^2}$$

are both distributed as χ_1^2. By the reproductive property of the chi-square distribution, we have

$$v_i = \frac{(b_i-\bar{b})^2 nN}{N-1} + \frac{(a_i^*-\bar{a})^2 nN}{N-1}$$

distributed as $\chi_2^2\sigma^2$, and this concludes the proof of (i).

If the v_i were distributed independently, the distribution of the largest of the v_i to the total of the v_i would be

given by (53). This distribution was determined by Fisher (1929), and a table of the 5% points of R' for values of N = 5(5)50 are given in his paper. Cochran (1941) extended the table to include values of N = 3(1)10. Since the v_i are not independently distributed, the distribution given by (53) is an approximation to the true distribution for N large. This approximation is good since $\rho_{v_i v_j} \longrightarrow 0$ as N increases. In order to verify this statement, we now derive the covariance of $V_i = v_i/\sigma^2$ and $V_j = v_j/\sigma^2$. Since

$$\mathrm{Cov}(V_i,V_j) = E(V_iV_j) - E(V_i)E(V_j),$$

we must calculate $E(V_iV_j)$, $E(V_i)$ and $E(V_j)$. First consider

$$E(V_iV_j) = E\left[\frac{(b_i-\bar{b})^2 + (a_i^*-\bar{a})^2}{\sigma^2(N-1)/nN} \cdot \frac{(b_j-\bar{b})^2 + (a_j^*-\bar{a})^2}{\sigma^2(N-1)/nN}\right]$$

$$= \frac{n^2N^2}{\sigma^4(N-1)^2} E[(b_i-\bar{b})^2(b_j-\bar{b})^2 + (b_i-\bar{b})^2(a_j^*-\bar{a})^2$$

$$+ (b_j-\bar{b})^2(a_i^*-\bar{a})^2 + (a_i^*-\bar{a})^2(a_j^*-\bar{a})^2].$$

Due to symmetry we need only evaluate $E(b_i-\bar{b})^2(b_j-\bar{b})^2$ and $E(b_i-\bar{b})^2(a_j^*-\bar{a})^2$. Letting $d_i = b_i-\beta$, $d_j = b_j-\beta$, we obtain

$$E[(b_i-\bar{b})^2(b_j-\bar{b})^2] = E[(d_i-\bar{d})^2(d_j-\bar{d})^2]$$

$$= E[d_i^2d_j^2 + d_j^2\bar{d}^2 - 2d_id_j^2\bar{d} + d_i^2\bar{d}^2 + \bar{d}^4$$

$$- 2d_i\bar{d}^3 - 2d_i^2d_j\bar{d} - 2d_j\bar{d}^3 + 4d_id_j\bar{d}^2]$$

$$= \frac{\sigma^4}{n^2} - \frac{2\sigma^4}{Nn^2} + \frac{3\sigma^4}{n^2N^2}.$$

Similarly, letting $c_j = a_j^* - \alpha$, we have

$$E[(b_i-\bar{b})^2(a_j^*-\bar{a})^2] = E[(d_i-\bar{d})^2(c_j-\bar{c})^2]$$

$$= E[d_i^2 - 2d_i\bar{d} + \bar{d}^2)(c_j^2 - 2c_j\bar{c} + \bar{c}^2]$$

$$= \left(\frac{\sigma^2}{n} - \frac{\sigma^2}{nN}\right)^2.$$

Therefore,

$$E(V_iV_j) = \frac{n^2N^2}{\sigma^4(N-1)^2}\left[\frac{2\sigma^4}{n^2} - \frac{4\sigma^4}{Nn^2} + \frac{6\sigma^4}{n^2N^2} + 2\left(\frac{\sigma^2}{n} + \frac{\sigma^2}{nN}\right)^2\right]$$

$$= \frac{4(N^2 - 2N + 2)}{(N-1)^2}.$$

Since $E(V_i) = E(V_j) = 2$ and $Var(V_i) = Var(V_j) = 4$, we have $Cov(V_i,V_j) = 4/(N-1)^2$ and $\rho_{V_iV_j} = 1/(N-1)^2$.

Corollary. If $\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 \neq n$, we have the following expression for R:

$$R = \frac{SSX(b_r-\bar{b})^2 + n(a_r^*-\bar{a})^2}{\sum_{i=1}^{N}[SSX(b_i-\bar{b})^2 + n(a_i^*-\bar{a})^2]},$$

where

$$SSX = \sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2.$$

Before concluding this section we propose an alternative procedure. We know that

$$T_j = \frac{v_j/2\sigma^2}{N(n-2)s^2/\sigma^2}, \text{ where } s^2 = \frac{1}{N(n-2)}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

possesses an F distribution with 2 and N(n-2) degrees of freedom. The outlier regression line may be detected by using the statistic, $\max_j T_j$, and the critical region is $\max_j T_j >$ constant. This critical region may be determined by using the joint probability density function of several correlated F ratios. Thus, let

$$P_0 = P(F_j < C,\ j = 1,2,\ldots,N)$$

$$= \int_0^C \int_0^C \cdots \int_0^C g(f_1,f_2,\ldots,f_N)\, df_1 df_2 \ldots df_N, \quad (54)$$

then

$$1 - P_0 = P(\text{At least one } F_j > C) = P(\max_j F_j > C).$$

$1 - P_0$ may be obtained by evaluation of the integral in (54). If the v_j in the numerator of T_j were independent we could use the results obtained by P. R. Krishnaiah and J. V. Armitage (1964). However, these results cannot be directly applied since the v_j are not independent. The derivation of the required multivariate F distribution will not be considered in this thesis.

C. A Test Criterion Based on Wilk's Statistic

In this section, the N lines described in Section B are represented by the following N points: $(\bar{y}_1,b_1)$, $(\bar{y}_2,b_2)$, ..., $(\bar{y}_N,b_N)$ where b_i is the least squares estimator of β and $\bar{y}_i = a_i^*$ is the least squares estimator of α^*. Thus we have a sample of size N from a two-dimensional normal distribution with mean vector (α^*,β) and variance-covariance matrix

$\begin{bmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2/SSX \end{bmatrix}$. It is assumed that α^*, β and the elements of the covariance matrix are unknown. Let

$$N\bar{Y} = \sum_{i=1}^{N} \bar{y}_i, \quad N\bar{b} = \sum_{i=1}^{N} b_i, \quad a_{11} = \sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2, \quad a_{22} = \sum_{i=1}^{N} (b_i - \bar{b})^2 \text{ and } a_{12} = a_{21} = \sum_{i=1}^{N} (b_i - \bar{b})(\bar{y}_i - \bar{Y}).$$

The sample may be represented by a cluster of N points in a two-dimensional euclidean space, R_2. Any two of these points together with the point $(\bar{Y},\bar{b})$ form a simplex. The sum of the squares of the volumes of all possible simplexes which can be so formed is shown by Wilks (1962) to be

$$(2!)^{-2} \left|a_{ij}\right|, \quad i,j = 1,2$$

where $\left|a_{ij}\right|$ is the determinant of the matrix $[a_{ij}]$, i.e.

$$\left|a_{ij}\right| = \sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2 \sum_{i=1}^{N} (b_i - \bar{b})^2 - \left[\sum_{i=1}^{N} (\bar{y}_i - \bar{Y})(b_i - \bar{b})\right]^2.$$

Wilks called $\left|a_{ij}\right|$ the internal scatter of the sample.

If the kth element of the sample is omitted, the internal scatter of the remaining N-1 points in the sample is denoted by $\left|a_{ijk}\right|$. Let

$$R_k = \frac{\left|a_{ijk}\right|}{\left|a_{ij}\right|}, \quad k = 1,2,\ldots,N.$$

The quantities $R_1, R_2, \ldots, R_N$ are called one-outlier scatter ratios of the sample. The criterion proposed for selecting and testing a single outlying point is $r_1 = \min_k (R_k)$. The

critical values of r_1 are in the left tail of the distribution. Due to mathematical difficulties, Wilks did not give the distribution of r_1; however, he did give upper bounds for $P(r_1 < r)$.

The ratio R_k applied to the present situation of outlying lines is

$$R_k = \frac{\sum_{i=1}^{N-1}(\bar{y}_i-\bar{Y}_k)^2 \sum_{i=1}^{N-1}(b_i-\bar{b}_k)^2 - [\sum_{i=1}^{N-1}(\bar{y}_i-\bar{Y}_k)(b_i-\bar{b}_k)]^2}{\sum_{i=1}^{N}(\bar{y}_i-\bar{Y})^2 \sum_{i=1}^{N}(b_i-\bar{b})^2 - [\sum_{i=1}^{N}(\bar{y}_i-\bar{Y})(b_i-\bar{b})]^2},$$

where

$$\bar{Y}_k = \frac{N\bar{Y}-\bar{y}_k}{N-1},$$

$$\bar{b}_k = \frac{N\bar{b}-b_k}{N-1},$$

and the terms in the numerator are summed over the deleted sample. Now

$$\sum_{i=1}^{N-1}(\bar{y}_i-\bar{Y}_k)^2 = \sum_{i=1}^{N}(\bar{y}_i-\bar{Y})^2 - \frac{N}{N-1}(\bar{y}_k-\bar{Y})^2,$$

$$\sum_{i=1}^{N-1}(b_i-\bar{b}_k)^2 = \sum_{i=1}^{N}(b_i-\bar{b})^2 - \frac{N}{N-1}(b_k-\bar{b})^2,$$

and

$$\sum_{i=1}^{N-1}(b_i-\bar{b}_k)(\bar{y}_i-\bar{Y}_k) = \sum_{i=1}^{N}(b_i-\bar{b})(\bar{y}_i-\bar{Y}) - \frac{N}{N-1}(b_k-\bar{b})(\bar{y}_k-\bar{Y}).$$

Then R_k may be written as

$$R_k = 1 - \frac{NA}{(N-1)B},$$

where

$$A = (\bar{y}_k - \bar{Y})^2 \sum_{i=1}^{N} (b_i - \bar{b})^2 + (b_k - \bar{b})^2 \sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2$$

$$- 2(b_k - \bar{b})(\bar{y}_k - \bar{Y}) \sum_{i=1}^{N} (b_i - \bar{b})(\bar{y}_i - \bar{Y}),$$

$$B = \sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2 \sum_{i=1}^{N} (b_i - \bar{b})^2 - \left[\sum_{i=1}^{N} (b_i - \bar{b})(\bar{y}_i - \bar{Y})\right]^2.$$

D. A Test Criterion Based on Siotani's Statistic

Siotani (1959) discussed the extreme value of the generalized distances, from the origin and the sample mean, of n points in a p-variate normal sample for the cases where the variance is known and where the variance is unknown. Tables giving the upper percentage points of the extreme deviate from the sample mean were determined for both cases.

Although our present problem involves only two regression coefficients, we first discuss the problem in general, and consider the case where we have N samples of size n involving k regression coefficients. In matrix notation, we have

$$Y_i = X_i \theta_i + e_i, \quad i = 1,2,\ldots,N.$$

where

$$Y_i = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad X_i = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \Theta_i = \begin{bmatrix} \Theta_1 \\ \Theta_2 \\ \cdot \\ \cdot \\ \cdot \\ \Theta_k \end{bmatrix}, \quad e_j = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix},$$

and the e_i's are $N(0,\sigma^2)$.

Then

$$\hat{\Theta}_i = (X_i'X_i)^{-1}X_i'Y_i,$$

and

$$\hat{\sigma}_i^2 = \frac{1}{n-k}(Y_i - X_i\hat{\Theta}_i)'(Y_i - X_i\hat{\Theta}_i).$$

Let us now assume that i) $S_i = X_i'X_i$ is independent of i, i.e. $X_i'X_i = S$ for all i, ii) σ_i common, and iii) Θ_i common. We then have

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_N \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_N \end{bmatrix} \Theta + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_N \end{bmatrix},$$

$$\hat{\Theta} = \left[\begin{bmatrix} X_1' & X_2' & \cdots & X_N' \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_N \end{bmatrix} \right]^{-1} \begin{bmatrix} X_1' & X_2' & \cdots & X_N' \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_N \end{bmatrix}$$

$$= \left[\sum_{i=1}^{N} X_i'X_i \right]^{-1} \sum_{i=1}^{N} X_i'Y_i$$

$$= \left[\sum_{i=1}^{N} S_i \right]^{-1} \sum_{i=1}^{N} X_i'Y_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} S^{-1} X_i' Y_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i,$$

and

$$\hat{\sigma}^2 = \frac{1}{Nn-k} \sum_{i=1}^{N} (Y_i - X_i\hat{\theta})'(Y_i - X_i\hat{\theta}).$$

These are the best estimators under the three assumptions mentioned previously. However, another unbiased estimator for σ^2 is

$$\hat{\hat{\sigma}}^2 = \frac{1}{N(n-k)} \sum_{i=1}^{N} (Y_i - X_i\hat{\theta}_i)'(Y_i - X_i\hat{\theta}_i).$$

$\hat{\theta}_i$, $i = 1,2,\ldots,N$, and $\hat{\theta}$ are unbiased estimators with variance-covariance matrices S_i^{-1}/σ^2 and S^{-1}/σ^2, respectively. We have made the assumption that the S_i, $i = 1,2,\ldots,N$, are independent of i, i.e. $S_i = S$ for all i. In this situation, the test procedure might be to test for

$$\frac{\max_i (\hat{\theta}_i - \hat{\theta}) \; S^{-1} \; (\hat{\theta}_i - \hat{\theta})}{\hat{\sigma}^2}.$$

However, the distribution of this statistic has not been found. Instead, we suggest using

$$\frac{\max_i (\hat{\theta}_i - \hat{\theta}) \; S^{-1} \; (\hat{\theta}_i - \hat{\theta})}{\hat{\hat{\sigma}}^2}.$$

The exact distribution of this statistic has also not been

studied so far. If N(n-k) is fairly large, however, $\hat{\sigma}^2$ could be taken as the true value of σ^2, and we could use with fairly good accuracy the upper percentage points obtained by Siotani (1959) for the case where the variance-covariance matrix is completely known. In general, when working with lines, the estimate of σ^2 may be quite good since the pooled degrees of freedom will be fairly large.

E. Empirical Comparison of the Criteria Discussed in Sections B, C and D

The model we have been using is specified concisely by the equation, $y = \alpha + \beta(x-\bar{x}) + e$, where y is any value of the dependent variable, x is fixed and e is a random variable drawn from $N(0,\sigma^2)$. In this section we generate ten lines from the model, $y = \alpha + \beta(x-\bar{x}) + e$, with $\alpha = 4$, $\beta = 0.5$ and e drawn from N(0,1), and one line from the same model with $\alpha = 6$, $\beta = 1$ and e drawn from N(0,1).

The procedure followed was to assign to x the values -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5. The value of $\alpha + \beta(x-\bar{x})$ for each of these x values was then determined. The variable part of y, i.e. e, was drawn at random by the procedure given in Snedecor (1956). We represent the eleven lines obtained by the points (4.3909, 0.4955), (4.1091, 0.5673), (3.6727, 0.4773), (4.0545, 0.5518), (3.5545, 0.3745), (4.4545, 0.2809), (3.8182, 0.5027), (3.4545, 0.4645), (4.0818, 0.4364), (4.2273, 0.5427), (6.1545, 1.1464), the first coordinate being $\bar{y}_i$ and

the second, b_i, $i = 1,2,\ldots,11$.

To apply the criterion given in Section B, we calculate $SSX(b_i-\bar{b})^2 + n(\bar{y}_i-\bar{Y})^2$ for each of the preceeding points. Dividing the largest of these by the total we obtain $R = 0.718$. The 5% critical point given by Fisher (1929) is approximately 0.445. Therefore, the test has located the outlying line.

Applying the criterion given in Section C to the same set of lines, we find that the internal scatter of the original set of eleven lines is 0.8365 and the scatter of the sample obtained by omitting the eleventh line is 0.7796. The ratio 0.7796/0.8365 = 0.093 is smaller than any of the ratios obtained by deleting any one of the first ten lines. According to Wilk's (1963) table, the upper bound at the 5% level is 0.260, i.e. P(ratio < 0.260) = 0.05. Since the observed ratio falls in the critical region we again reject the eleventh line.

In applying the criterion given in Section D, we may use the same values given previously for the slope and intercept because in this example $\bar{x} = 0$. We have

$$\frac{\max_i (\hat{\theta}_i - \hat{\theta})' S^{-1} (\hat{\theta}_i - \hat{\theta})}{\hat{\sigma}^2} = 96.5,$$

where the maximum is obtained by using the eleventh line. Using Siotani's table we find that the upper bound at the 5% level lies between 9.48 and 9.99. Since the observed value falls in the critical region, we again reject the eleventh

line.

F. Outlier Tests Assuming Available A Priori Information Sufficient to Identify a Suspected Outlier

1. Construction of the U statistic

In Section E of Chapter IV we let $Y_{oj} = a_o + b_o x_{oj}$ represent the suspected outlier line and $Y_{aj} = a_a + b_a x_{aj}$, which was obtained by using all the data except that pertaining to the suspected outlier line, represent the estimated average regression line. We now combine the data used to determine these two lines and obtain a third regression line, say $Y_{cj} = a_c + b_c x_{cj}$, for the combined set of observations. The data needed are

Sample size n_o+n_a

Mean values $\bar{x}_c = (n_o\bar{x}_o + n_a\bar{x}_a)/(n_o+n_a)$

$\bar{y}_c = (n_o\bar{y}_o + n_a\bar{y}_a)/(n_o+n_a)$

Corrected sums of squares and products $S = S_o + S_a + n_o n_a(\bar{x}_o-\bar{x}_a)^2/(n_o+n_a)$

$Q = Q_o + Q_a + n_o n_a(\bar{x}_o-\bar{x}_a)(\bar{y}_o-\bar{y}_a)/(n_o+n_a)$.

Then

$$R_1 = \sum_{j=1}^{n_o}(y_{oj}-\bar{y}_o)^2 + \sum_{j=1}^{n_a}(y_{aj}-\bar{y}_a)^2 - b_c Q + \frac{n_o\, n_a}{n_o+n_a}(\bar{y}_o-\bar{y}_a)^2$$

has n_o+n_a-2 degrees of freedom.

The above information is summarized in the following analysis of variance table.

Table 9. Analysis of variance

Source	df	SS	MS
Deviations from hypothesis	2	R_1-R_0	$(R_1-R_0)/2$
Separate regressions	n_o+n_a-4	R_0	$R_0/(n_o+n_a-4)$
Common regressions	n_o+n_a-2	R_1	

The criterion used for rejecting the hypothesis that a line is not an outlier line is the following: We test the hypothesis H_0: $\alpha_o = \alpha_a$, $\beta_o = \beta_a$ versus H_1: $\alpha_o \neq \alpha_a$, $\beta_o \neq \beta_a$. If the hypothesis is not rejected, we conclude that we have no reason to say that the line is an outlier. In order to test the hypothesis, we use the statistic

$$U = \frac{(R_1 - R_0)/2}{R_0/(n_o+n_a-4)}.$$

In the next section we show that if the null hypothesis be true, U is distributed as an F with 2 and n_o+n_a-4 degrees of freedom.

2. Derivation of the distribution of U

Theorem 4. Let U be a random variable of the form

$$U = \frac{(R_1 - R_0)/2}{R_0/(n_o+n_a-4)},$$

where R_1-R_0 and R_0 are random variables defined previously. Then, under the null hypothesis, U is distributed as an F with with 2 and n_o+n_a-4 degrees of freedom.

proof: In Section E of Chapter IV, we showed that R_o/σ^2

is always distributed as a χ^2 with n_o+n_a-4 degrees of freedom, and hence that $E[R_0/(n_o+n_a-4)] = \sigma^2$. Now consider the numerator of U. We have

$$\frac{R_1 - R_0}{\sigma^2} = \frac{K}{\sigma^2 S(S_o+S_a)}[(S_o+S_a)(\bar{y}_o-\bar{y}_a) - (Q_o+Q_a)(\bar{x}_o-\bar{x}_a)]^2$$

$$+\frac{1}{\sigma^2 S_o S_a (S_o+S_a)}(Q_a S_o - Q_o S_a)^2,$$

where $K = n_o n_a/(n_o+n_a)$, or

$$\frac{R_1 - R_0}{\sigma^2} = Z^2 + T^2,$$

where

$$Z = \frac{\sqrt{K}}{\sqrt{\sigma^2 S(S_o+S_a)}}[(S_o+S_a)(\bar{y}_o-\bar{y}_a) - (Q_o+Q_a)(\bar{x}_o-\bar{x}_a)],$$

$$T = \frac{1}{\sqrt{\sigma^2 S_o S_a (S_o+S_a)}}(Q_a S_o - Q_o S_a).$$

Since T and Z are linear combinations of normally distributed variables, they are themselves normally distributed. We see that

$$E(Z) = \frac{\sqrt{K}}{\sqrt{\sigma^2 S(S_o+S_a)}}[(S_o+S_a)(\alpha_o-\alpha_a+\beta_o\bar{x}_o-\beta_a\bar{x}_a) - (\bar{x}_o-\bar{x}_a)$$

$$\cdot(\beta_o S_o+\beta_a S_a)],$$

and

$$E(Z^2) = \frac{K}{\sigma^2 S(S_o+S_a)}\Big\{(S_o+S_a)^2[\sigma^2/K + (\alpha_o-\alpha_a+\beta_o\bar{x}_o-\beta_a\bar{x}_a)^2]$$
$$+ (\bar{x}_o-\bar{x}_a)^2(\beta_o^2S_o^2+\sigma^2S_o+\beta_a^2S_a^2+\sigma^2S_a+2\beta_o\beta_aS_oS_a)$$
$$-2(S_o+S_a)(\bar{x}_o-\bar{x}_a)(\alpha_o\beta_oS_o+\beta_o^2\bar{x}_oS_o+\alpha_o\beta_aS_a+\beta_o\beta_a\bar{x}_oS_a$$
$$-\alpha_a\beta_oS_o-\beta_o\beta_aS_o\bar{x}_a-\alpha_a\beta_aS_a-\beta_a^2\bar{x}_aS_a)\Big\}.$$

Under the null hypothesis, $E(Z) = 0$ and $E(Z^2) = 1$. Hence Z is distributed as a standard normal and Z^2 is distributed as a χ^2 with one degree of freedom. Now consider the mean and variance of T. We have

$$E(T) = \frac{S_oS_a}{\sqrt{\sigma^2S_oS_a(S_o+S_a)}}(\beta_a-\beta_o)$$

and

$$E(T^2) = \frac{S_oS_a(\beta_o-\beta_a)^2 + \sigma^2(S_o+S_a)}{\sigma^2(S_o+S_a)}.$$

Under the null hypothesis, $E(T) = 0$ and $E(T^2) = 1$. Thus T is distributed as a standard normal and T^2 is distributed as a χ^2 with one degree of freedom. Consequently, $(R_1 - R_0)/\sigma^2$ is distributed as a χ^2 with two degrees of freedom when the null hypothesis is true.

3. Distributional structure of U under the alternative hypothesis

Under the alternative hypothesis the numerator of the U statistic will not be distributed as a χ^2 but rather as a non-

central χ^2 with non-centrality parameter given by

$$\lambda = \frac{1}{2S(S_o+S_a)\sigma^2}\Big\{K(S_o+S_a)^2(\alpha_o-\alpha_a)^2 + 2K(S_o+S_a)(S_a\bar{x}_o+\bar{x}_aS_o)$$

$$(\alpha_o-\alpha_a)(\beta_o-\beta_a) + [SS_oS_a + K(S_a\bar{x}_o+S_o\bar{x}_a)^2](\beta_o-\beta_a)^2\Big\}.$$

To verify this, we note that under the alternative hypothesis, T is distributed as a normal with mean E(T) and variance 1, and Z is distributed as a normal with mean E(Z) and variance 1. Hence, T^2 is distributed as a non-central χ^2 with one degree of freedom and non-centrality parameter $\lambda_1 = E^2(T)/2$; Z^2 is distributed as a non-central χ^2 with one degree of freedom and non-centrality parameter $\lambda_2 = E^2(Z)/2$. Therefore, the numerator of U is distributed as a non-central χ^2 with two degrees of freedom and non-centrality parameter given by $\lambda_1+\lambda_2 = \lambda$.

The denominator of U is always distributed as a central χ^2. Thus, under the alternative hypothesis, U is distributed as a non-central F with 2 and n_o+n_a-4 degrees of freedom and non-centrality parameter λ.

The power of the test for a special case is given in the next section.

4. Special case

We assume that the same set of x's is used for each of the N lines, that $\bar{x}_o = \bar{x}_a = 0$, $S_a = (N-1)S_o$, $n_a = (N-1)n_o$, $S = NS_o$, $S_o = n_o$, and $S_a = n_a$. In this special situation, the

non-centrality parameter is $[n_o(N-1)(A^2+B^2)]/2N$, where $A = (\alpha_o-\alpha_a)/\sigma$ and $B = (\beta_o-\beta_a)/\sigma$.

Table 10. Power of the test when $N = 6$ and $n_o = 5$

|A|	|B| = 0	1	2	3
0	0.000	0.186	0.787	0.984
1	0.186	0.419	0.886	0.990
2	0.787	0.886	0.977	
3	0.984	0.990		

5. Construction of the statistic U'

A different approach to the problem of testing H_0 is the following: First, find the $100(1-\alpha_1)\%$ joint confidence region for α_a and β_a, the regression coefficients for the average line. This confidence region is the interior of an ellipse whose equation is

$$n_a(\alpha-a_a)^2 + 2n_a\bar{x}_a(\alpha-a_a)(\beta-b_a) + \sum_{j=1}^{n_a} x_{aj}^2(\beta-b_a)^2 = 2F_{2,n_a-2}\, s_a^2, \qquad (55)$$

where

$s_a^2 = \sum_{j=1}^{n_a} (y_{aj}-Y_{aj})^2/(n_a-2)$, and a_a and b_a are the point estimates of α_a and β_a, respectively.

Next, find the $100(1-\alpha_1)\%$ joint confidence region for α_o and β_o, the regression coefficients for the suspected outlier line. This confidence region is also the interior of an

ellipse and its equation is

$$n_o(\alpha-a_o)^2 + 2n_o\bar{x}_o(\alpha-a_o)(\beta-b_o) + \sum_{j=1}^{n_o} x_{oj}^2(\beta-b_o)^2 = 2F_{2,n_o-2}\, s_o^2, \qquad (56)$$

where $s_o^2 = \sum_{j=1}^{n_o} (y_{oj}-Y_{oj})^2/(n_o-2)$, and a_o and b_o are the point estimates of α_o and β_o, respectively.

We wish to test the hypothesis H_0: $\alpha_o = \alpha_a$, $\beta_o = \beta_a$, i.e. we wish to test the hypothesis that a particular line is not an outlier. The criterion considered here for the rejection of this hypothesis is the following: If the two confidence regions, given by (55) and (56), fail to overlap, we reject the hypothesis.

Let us now assume that the same set of x's is selected for each of the N lines such that $\bar{x}_i = 0$ and $\sum_{j=1}^{n} x_{ij}^2 = n$. Then (55) and (56) become

$$(N-1)n\,(\alpha-a_a)^2 + (N-1)n(\beta-b_a)^2 = 2F_{2,n_a-2}\, s_a^2$$

and

$$n(\alpha-a_o)^2 + n(\beta-b_o)^2 = 2F_{2,n-2}\, s_o^2.$$

We now have two circles instead of two ellipses. These two circles will fail to intersect if the distance between their centers exceeds the sum of their radii. That is, if

$$[(a_o-a_a)^2 + (b_o-b_a)^2]^{1/2} \geq [2F_{2,n_a-2}/(N-1)n]^{1/2}\, s_a + [2F_{2,n-2}/n]^{1/2}\, s_o,$$

or $U' \geq 1$, where

$$U' = \frac{(a_o-a_a)^2 + (b_o-b_a)^2}{(\phi_1 s_a + \phi_2 s_o)^2},$$

$$\phi_1 = [2F_{2,n_a-2}/n(N-1)]^{1/2},$$

$$\phi_2 = [2F_{2,n-2}/n]^{1/2}.$$

Therefore, the test procedure is to calculate U' and reject the hypothesis H_0 if $U' \geq 1$ and accept otherwise.

6. Distributional structure of the U' statistic under the null hypothesis

Under the null hypothesis the expression

$$\frac{C^2 + D^2}{E},$$

where

$$C = \frac{(a_o-a_a) - (\alpha_o-\alpha_a)}{\sqrt{\sigma^2 N/n(N-1)}},$$

$$D = \frac{(b_o-b_a) - (\beta_o-\beta_a)}{\sqrt{\sigma^2 N/n(N-1)}},$$

and

$$E = \frac{(\phi_1 s_a + \phi_2 s_o)^2}{\sqrt{\sigma^2 N/n(N-1)}},$$

reduces to the U' statistic. Since (a_o-a_a) is distributed as a normal variate with mean $(\alpha_o-\alpha_a)$ and variance $\sigma^2 N/n(N-1)$, and (b_o-b_a) is distributed as a normal variate with mean $(\beta_o-\beta_a)$ and variance $\sigma^2 N/n(N-1)$, we know that both C and D are dis-

tributed as standard normal variates. Therefore, under the null hypothesis C^2+D^2 is distributed as a χ^2 with 2 degrees of freedom. Now consider the quantity E. This may be written as

$$(\delta_1 \chi_{n_a-2} + \delta_2 \chi_{n-2})^2,$$

where

$$\delta_1 = [2F_{2,n_a-2}/N(n_a-2)]^{1/2},$$

$$\delta_2 = [2(N-1)F_{2,n-2}/N(n-2)]^{1/2}.$$

Hence, under the null hypothesis, U' is distributed as

$$\frac{\chi_2^2}{(\delta_1 \chi_{n_a-2} + \delta_2 \chi_{n-2})^2}.$$

7. Approximate distribution of U'

We can approximate the quantity $(\delta_1 \chi_{n_a-2} + \delta_2 \chi_{n-2})$ by using the procedure outlined in Section E of Chapter IV. In this case we have

$$H_1 = \delta_1^2(n_a-2) + \delta_2^2(n-2) + 2\delta_1\delta_2 b_1 b_2,$$

$$H_2 = \delta_1^4 n_a(n_a-2) + \delta_2^4 n(n-2) + 6\delta_1^2\delta_2^2(n-2)(n_a-2) +$$

$$+ 4\delta_1^3\delta_2 b_1 b_2(n_a-1) + 4\delta_1\delta_2^3 b_1 b_2(n-1),$$

where

$$b_1 = 2^{1/2}\,\Gamma[(n_a-1)/2] / \Gamma[(n_a-2)/2] \text{ and}$$

$$b_2 = 2^{1/2}\,\Gamma[(n-1)/2] / \Gamma[(n-2)/2].$$

The approximate distribution of U' under the null hypothesis is given by

$F_{2,\nu}/\gamma$, where $\gamma = (H_2 - H_1^2)/2H_1$ and $\nu = 2H_1^2/(H_2 - H_1^2)$.

8. Size of the test using the approximate distribution

The probability of a type I error or the size of the critical region is given by

$$P(F_{2,\nu}/\gamma \geq 1) = P(F_{2,\nu} \geq \gamma).$$

The size of the test depends upon (i) the value of α_1 used in determining the confidence regions, (ii) the number of lines in the experiment and (iii) the number of points used to determine each line.

No attempt is made in this thesis to study the power of the U' test.

VI. SUMMARY

The general purpose in making outlier tests depends upon the aim of the experiment. The aim may be (1) to identify possible outliers either as an end in itself or in order to investigate the conditions which may have led to this outlying observation; (2) to estimate population parameters or to test hypotheses. If the latter is the aim of the experiment then the outlier test is made to determine which sample values should be used to make these subsequent population inferences and this outlier test should be taken into account when making these inferences. These two purposes are discussed in a paper by Quesenberry and David (1961).

This thesis represents the first attempt to obtain a statistical outlier methodology when (2) is the aim of the experiment. We consider the problems of estimation and hypothesis testing subsequent to a preliminary test for a univariate statistical outlier. We investigate these problems (a) when the scientist has performed the preliminary test for an outlying observation assuming no a priori information, and (b) when the scientist has performed the preliminary test for an outlying observation assuming a priori information sufficient to identify a suspected outlier. In both situations we simplify the problem by considering only the case where one observation is suspect. Formulae for the bias, mean square error and power are derived and some numerical values obtained. These values

should not be used as absolute estimates since they are dependent on population parameters.

We also consider the problem of line outliers and present several approximate and hindsight statistics. In this connection we note the following. Suppose one is interested in pruning the data in order to obtain a more accurate subsequent inference and does not take into account the fact that a preliminary test was made. Then one might make inaccurate inferences even though the proper outlier test (i.e. one based on the fact that a particular line is tested) is used. It may even be possible that the error in the inferences will be larger than if one used an approximate outlier test, but took this test into account when making further inferences. In other words, if one has a choice of ignoring the effect of a preliminary test on subsequent estimation, or using an inaccurate preliminary test but considering its effect on subsequent estimation, then the latter procedure might be the better one if the inaccurate preliminary test is not too bad an approximation.

Concerning work carried out on approximate tests for line outliers, it is difficult to judge which method of solution is to be preferred. The same difficulty arises when we consider the many tests proposed for univariate outliers. Until a study is made of the power of the tests it is impossible to choose one and say that it is best. It is for this reason that so many criteria are now in use.

VII. LITERATURE CITED

Airy, G. B. 1856. Letter from Professor Airy, Astronomer Royal, to the editor. Astronomical Journal. 4:137-138. Original not available; cited in Rider, P. R. 1933. Criteria for rejection of observations. Washington University Studies-New Series, Science and Technology No. 8: 21.

Anscombe, F. J. 1960. Rejection of outliers. Technometrics. 2:123-147.

Bancroft, T. A. 1964. Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance. Biometrics. 20:427-442.

Bennett, B. M. 1952. Estimation of means on the basis of preliminary tests of significance. Ann. Inst. of Stat. Math. 4:31-43.

Bennett, B. M. 1956. On the use of preliminary tests in certain statistical procedures. Ann. Inst. of Stat. Math. 8:45-57.

Chauvenet, W. 1876. A manuel of spherical and practical astronomy. (University edition) Vol. 2. Philadelphia: Lippincott; and London: Trubner. 5th edition. Original not available; cited in Rider, P. R. 1933. Criteria for rejection of observations. Washington University Studies -New Series, Science and Technology No. 8:8.

Cochran, W. G. 1941. The distribution of the largest of a set of estimated variances as a fraction of their total. Annals of Eugenics. 2:47-52.

Czuber, E. 1891. Theorie der Beobachtungsfehler. Leipzig, Teuber. Original not available; cited in Rider, P. R. 1933. Criteria for rejection of observations. Washington University Studies-New Series, Science and Technology No. 8:10.

Dixon, W. J. 1950. Analysis of extreme values. Ann. Math. Stat. 21:488-506.

Dixon, W. J. 1953. Processing data for outliers. Biometrics. 9:47.

Dixon, W. J. and Massey, F. J. 1951. Introduction to statistical analysis. 1st ed. New York, N. Y., McGraw-Hill Book Co., Inc.

Fisher, R. A. 1929. Tests of significance in harmonic analysis. Proc. Roy. Soc. Series A, 125:54-59.

Greenwood, J. R. and Hartley, H. O. 1962. Guide to tables in mathematical statistics. Princeton, N. J., Princeton University Press.

Grubbs, F. E. 1950. Sample criteria for testing outlying observations. Ann. Math. Stat. 21:27-58.

Irwin, J. O. 1925. On a criterion for the rejection of outlying observations. Biometrika. 17:238-250.

Kenney, J. F. and Keeping, E. S. 1951. Mathematics of Statistics. Part 2. 2nd ed. New York, N. Y., D. Van Nostrand Co., Inc.

Kitagawa, T. 1950. Successive process of statistical inference. Mem. Fac. Sci. Kyusyu Univ. Series A, 5:139-180.

Krishnaiah, P. R. and Armitage, J. V. 1965. Tables for the distribution of the maximum of correlated chi-square variates with one degree of freedom. ARL 65- . Aero- Wright-Patterson Air Force Base, Ohio, Aerospace Research Laboratories. (In preparation)

Krishnaiah, P. R. and Armitage, J. V. 1964. Probability integrals of the multivariate F distribution, with tables for special cases. Unpublished paper presented at the Central Regional meeting of the Inst. of Math. Stat., Chicago, Ill., Dec. 1964. Mimeo. Wright-Patterson Air Force Base, Ohio, Aerospace Research Laboratories.

Kudô, A. 1956. On the testing of outlying observations. Sankhya. 17:67-76.

Li, J. 1964. Statistical inference. Ann Arbor, Mich., Edwards Brothers, Inc.

McCullough, R. S. 1961. Testing equality of means under variance heterogeneity. Unpublished Ph.D. thesis. Ames, Iowa, Library, Iowa State University of Science and Technology.

McKay, A. T. 1935. The distribution of the difference between the extreme observation and the sample mean in samples of n from a normal universe. Biometrika. 27:466-471.

Mosteller, F. 1948. A k-sample slippage test for an extreme population. Ann. Math. Stat. 19:58-65.

Nair, K. R. 1948. The distribution of the extreme deviate from the sample mean and its studentized form. Biometrika. 35:118-144.

Patnaik, P. B. 1949. The non-central χ^2 and F distributions and their applications. Biometrika. 36:202-232.

Pearson, E. S. 1932. The percentage limits for the distribution of range in samples from a normal population. Biometrika. 24:404-417.

Pearson, E. S. and Hartley, H. O. 1943. Tables of the probability integral of the studentized range. Biometrika. 33:89-99.

Pearson, E. S. and Hartley, H. O. 1942. The probability integral of the range in samples of n observations from a normal population. Biometrika. 32:301-310.

Pearson, E. S. and Sekar, C. C. 1936. The efficiency of statistical tools and a criterion for the rejection of outlying observations. Biometrika. 28:308-320.

Pearson, K. 1931. Tables for statisticians and biometricians. Part 2. Cambridge, Cambridge University Press.

Pierce, B. 1852. Criterion for the rejection of doubtful observations. Astronomical Journal. 2:161-163. Original not available; cited in Rider, P. R. 1933. Criteria for rejection of observations. Washington University Studies -New Series, Science and Technology No. 8:5.

Pillai, K. C. S. 1959. Upper percentage points of the extreme studentized deviate from the sample mean. Biometrika. 46:473-474.

Quesenberry, C. P. and David, H. A. 1961. Some tests for outliers. Biometrika. 48:379-390.

Rao, C. R. 1952. Advanced statistical methods in biometric research. New York, N. Y., John Wiley and Sons, Inc.

Rider, P. R. 1933. Criteria for rejection of observations. Washington University Studies-New Series, Science and Technology No. 8.

Siotani, M. 1959. The extreme value of the generalized distances of the individual points in the multivariate normal sample. Ann. Inst. of Stat. Math. 10:183-206.

Snedecor, G. W. 1956. Statistical methods. 5th ed. Ames, Iowa, Iowa State University Press.

Stewart, R. M. 1920. Peirce's criterion. Popular Astronomy. 28:2-3. Original not available; cited in Rider, P. R. 1933. Criteria for rejection of observations. Washington University Studies-New Series, Science and Technology No. 8:17.

Stone, E. J. 1867. On the rejection of discordant observations. Monthly Notices of the Royal Astronomical Society. 28:165-168. Original not available; cited in Rider, P. R. 1933. Criteria for rejection of observations. Washington University Studies-New Series, Science and Technology No. 8:9.

Tang, P. C. 1938. The power function of the analysis of variance tests with tables and illustrations of their use. Stat. Res. Mem. 2:126-149.

Thompson, W. R. 1935. On a criterion for the rejection of observations and the distribution of the ratio of the deviation to the sample standard deviation. Ann. Math. Stat. 6:214-219.

Tippett, L. H. C. 1925. The extreme individuals and the range of samples taken from a normal population. Biometrika. 17:151-164.

Tukey, J. W. 1949. Comparing individual means in the analysis of variance. Biometrics. 5:99-114.

Wilks, S. S. 1962. Mathematical statistics. New York, N. Y., John Wiley and Sons, Inc.

Wilks, S. S. 1963. Multivariate statistical outliers. Sankhya. 25:407-426.

Wright, T. W. 1884. A treatise on the adjustment of observations by the method of least squares. New York, N. Y., D. Van Nostrand Co., Inc.

VIII. ACKNOWLEDGEMENTS

The author wishes to acknowledge her indebtedness to Dr. T. A. Bancroft for suggesting the problem and wishes to express her sincere appreciation to him for his interest, assistance and encouragement during the preparation of this thesis. She is also grateful to Dr. B. K. Kale and Dr. A. Kudô for some helpful suggestions with regard to some of the approaches used. Part of the research on this thesis was sponsored by NSF GP 1855 and NSF GP 274.